

R E P O R T R E S U M E S

ED 002 839

24

MEASUREMENT OF THE EDUCABILITY OF SEVERELY MENTALLY RETARDED CHILDREN.

BY- GRACE, ALONZO AND OTHERS
NEW YORK UNIV., SCH. OF EDUCATION

REPORT NUMBER CRP-332

PUB DATE 59

CONTRACT OEC-SAE-7783

EDRS PRICE MF-\$0.18 HC-\$4.60 115P.

DESCRIPTORS- *MENTAL RETARDATION, RETARDED CHILDREN, *EDUCABLE MENTALLY HANDICAPPED, *PSYCHOMETRICS, LEARNING PROCESSES, *ABILITY IDENTIFICATION, ACADEMIC ABILITY, RETENTION, TRAINING, TRANSFER OF TRAINING, *MENTAL TESTS, DIAGNOSTIC TESTS (EDUCATION), NEW YORK CITY, NEW YORK

THE PURPOSE OF THIS STUDY WAS TO INVESTIGATE A TESTING METHOD FOR OBTAINING OBJECTIVE, QUANTIFIED, PREDICTIVE MEASURES OF EDUCABILITY FOR MENTAL RETARDATE. THE METHOD WAS DESIGNED TO MEASURE ABILITIES ON WHICH EDUCATION DEPENDS--NAMELY, THE ABILITY TO LEARN UNDER TRAINING, TO TRANSFER THE TRAINING, AND TO RETAIN THE LEARNING. TWO SAMPLES WERE USED--A PILOT GROUP OF 115 AND A FINAL GROUP OF 114. SUBJECTS IN THE FINAL GROUP HAD A MEAN IQ OF 36 AND AN AVERAGE AGE OF 95 MONTHS. THE FINAL TEST BATTERY WAS ADMINISTERED INDIVIDUALLY TO TEACH EACH SUBJECT AND INCLUDED A TEST OF IMITATIVE ABILITIES AND MEMORY, A SIZE DISCRIMINATION TEST, A SHAPE DISCRIMINATION TEST, AND TWO TESTS OF BRIGHTNESS DISCRIMINATION. ALL OF THE SUBJECTS RECEIVED THE FIRST FOUR TEST ADMINISTRATIONS WHICH YIELDED PRETRAINING, POST-TRAINING, TRANSFER, AND RETENTION SCORES RESPECTIVELY. OF THE TOTAL SAMPLE, 102 TOOK THE FIFTH TEST SERIES, BASED ON RETENTION, AND 36 WERE AVAILABLE FOR THE LAST TESTS, REPETITIONS OF THE FIRST 4 TESTS. INTELLIGENCE EXERTED A GREATER SIGNIFICANT EFFECT ON THE SCORES OF EVERY TEST THAN THE FACTORS OF AGE, TIME SPENT IN SCHOOL, SEX, OR MEDICAL DIAGNOSTIC CATEGORY. THE RELATIONSHIP BETWEEN LEARNING PERFORMANCE AND INTELLIGENCE LEVEL WAS SIGNIFICANTLY GREATER ON EVERY TEST THAN THAT BETWEEN INTELLIGENCE LEVEL AND INITIAL PERFORMANCE. THE ANCILLARY FINDINGS SUGGEST THE APPLICABILITY OF THE MATERIALS AND TRAINING METHODS DEVELOPED FOR THE TESTS TO CURRICULAR MATERIALS AND CLASSROOM METHODS, AND THE SUITABILITY OF TESTING USED FOR OTHER GROUPS OF CHILDREN. (JH)

ED002839

U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
Office of Education

This document has been reproduced exactly as received from the person or organization originating it. Points of view or opinions stated do not necessarily represent official Office of Education position or policy.

Measurement of the Educability of Severely Mentally Retarded Children

School of Education
New York University

Associate Dean Alonzo Grace,
in charge of research at the School of Education,
New York University

Adjunct Associate Professor Howard Newburger,
Project Director

Associate Research Scientist Helen Schucman,
Principal Investigator

The research reported herein was performed
pursuant to Contract #SAE-7783,
with the United States Office of Education,
Department of Health, Education, and Welfare

1959

Acknowledgments

Gratitude is due to the following members of the faculty of the School of Education of New York University: Professors Kenneth Benson, Roscoe Brown, Irene Cypher, Milton Gabrielsen, John Kinsella, Madeline Levine, and Elizabeth Trumbower,

to: Professor Christopher De Prospe, past President of the American Association on Mental Deficiency,
Professor Harold Michal-Smith, director of the Division of Psychology, New York Medical College, Flower and Fifth Avenue Hospitals,
James Toolan, M.D., Department of Psychiatry, New York University College of Medicine and Bellevue Hospital

and to: The National Association for Retarded Children, The New York Association for the Help of Retarded Children, Parents and Friends of Retarded Children of Bridgeport, Conn., The Nassau County Association for the Help of Retarded Children, Long Island, New York, The Rugby School, Brooklyn, New York, the Shield of David Institute for Retarded Children, New York City, the Association for Retarded Children, Orange, N. J., and the Mental Retardation Clinic at Flower and Fifth Avenue Hospitals of New York.

Table of Contents

<u>Chapter</u>	<u>Page</u>
I The Problem.....	1
II Method.....	4
III Results.....	22
IV Discussion.....	49
V Summary and Conclusions.....	64
Bibliography.....	70
Appendix I: Materials and Procedures for Administering the Tests.....	72
Appendix II: The Experimental Test Battery.....	96

Index of Tables

<u>Table</u>	<u>Page</u>
1. Product-Moment Correlation Coefficients Obtained by Test-Retest Comparisons.....	23
2. Multiple Correlation Coefficients Between the Five Independent Variables and Administrations of Each Test.....	30
3. Levels of Time in School Used in Analyses of Covariance.	32
4. Levels of Time in School Used in Analyses of Variance...	33
5. Levels of Age Used in Analyses of Covariance.....	33
6. Levels of Age Used in Analyses of Variance.....	34
7. Levels of Intelligence Used in Analyses of Covariance....	37
8. Levels of Intelligence Used in Analyses of Variance.....	41
9. Product-Moment Correlation Coefficients Between Estimated IQs, Stanford-Binet IQs, Teachers' Ratings, and Vineland Social Quotients.....	44
10. Product-Moment Correlation Coefficients Between Test Scores and Estimated IQs, Teachers' Ratings, and Vineland Social Quotients.....	46

Index of Figures

<u>Figures</u>	<u>Page</u>
1. Differences in Mean Scores for Each Administration of Each Test.....	27
2. Mean Scores for Various Administrations.....	29
3. Mean Scores of Analyses of Variance for Levels of Intelligence.....	40
4. The Two-Buttons Test.....	73
5. The Buzz Test.....	78
6. The Music Box Test.....	82
7. The Light Test.....	87
8. The Three-Buttons Test.....	91

I

THE PROBLEM

This chapter discusses the problem which was investigated, the method by which it was investigated, and the rationale of the method.

The Problem

The purpose of this study was to investigate a method for obtaining an objective, quantified, predictive measure of educability for the severely mentally retarded child. With increasing attention focussed on the education of these children, the need for suitable methods for evaluating their abilities has become acute. The education of the retarded child must concentrate on fostering his abilities in whatever direction and to whatever degree is possible.⁽¹⁾ His limited abilities must be assessed, and attention directed to developing the most profitable ones.⁽²⁾

Unfortunately, it is precisely these children whose abilities are among the most difficult to evaluate by the usual testing procedures. On standard tests, the child may be heavily penalized by verbal, motor, sensory, and experiential handicaps. He may also be unable to cope with the contents of the tests. Tests with which a child cannot cope will hardly serve to identify his assets. No curriculum can be built on what a child can not do.

It has been said that "the ideal test in the study of mental deficiency would be one which investigated the ability to learn."⁽³⁾ This research was undertaken to investigate the learning abilities on which education depends; namely, the ability to learn under training, to transfer the training, and to retain the learning, and to use the results as the basis for evaluating educability in severely retarded children.

Method

A battery of tests was administered to a group of severely retarded children. The children were tested individually. The subject was first given the tests without training. He was then trained to the correct responses according to standardized training procedures, and retested to permit measurement of posttraining learning gains. Different forms of the tests were used to evaluate transfer. Retention scores were obtained by retesting the subject at various later dates.

Rationale of the Method

Educability is an inference. It can be evaluated only in terms of what it enables a child to do. Logically, then, if we wish to measure educability, the first step is to determine what a child must be able to do in order to be educable.

Educability can virtually be defined in terms of a child's abilities to learn in response to instruction, to transfer the training, and to retain the learning. We establish schools to secure desired learning outcomes, and "it might almost be said that if there is to be education there must be transfer."⁽⁴⁾ If we wish to measure a child's abilities in these areas, we must place him in learning situations which are appropriate to his abilities, and which allow him to respond within his limitations.

The specific aims of this study were to develop a series of suitable learning tests for the severely mentally retarded child; to obtain samples of his performance in learning areas vital to education; and to relate his performance in these test situations to significant aspects of his life behavior, including variables which reflect classroom learning.

References

1. Porteus, S.D., Studies in Mental Deviations, The Vineland Training School at Vineland, New Jersey, 1923, p. v.
2. McLachlan, D.G., "Emotional Aspects of the Backward Child," Amer. J. of Ment Defic., Vol. 60, 1955, p. 330.
3. Penrose, L.S., Mental Defect, New York: Farrar and Rinehart, 1934, P. 49.
4. Pressey, S.L. and Robinson, F.P., Psychology and the New Education, New York: Harper and Bros., 1944, P. 573.

II

METHOD

This chapter describes the preliminary experimentation, the subjects, the final test battery, and general testing and training procedures.

The Preliminary Experimentation

The major steps in the preliminary experimentation were determining the plan of the experimental battery, constructing the preliminary tests, and conducting the pilot studies.

The Plan of the Experimental Battery

A series of short tests was anticipated, to hold the interest of the children, and to allow the battery to be administered in a number of sessions if necessary. Each test was to include items of increasing difficulty. Though short tests with relatively few items were essential, it was not likely that such a test would be equally suitable for purposes of discrimination throughout the range of abilities studied. It was therefore decided that the battery as a whole should assume this function. For this reason, tests characterized by varying degrees of difficulty were sought, the degree of difficulty to be determined by the number of items which were passed by the group as a whole.

The Construction of the Tests

Before the preliminary battery was constructed, the following criteria were set up to serve as guides in the construction and development of the tests:

1. Tests should be within the abilities of the subjects.

2. Tests should depend chiefly on abilities sufficiently well developed at the age levels of the subjects.
3. Tests should be suitable for brain-injured subjects.
4. Tests should be interesting to the subjects.
5. Tests should be short enough to hold the interest of the subjects.
6. Performance should depend minimally on past experience.
7. Tests should not depend on verbal instructions.
8. No verbal or manipulative responses should be required.
9. Tests should allow for possible difficulties in color perception.

In advance of the pilot studies, preliminary attempts were made to meet these criteria. The literature was reviewed for areas likely to be in accord with Criterion 1. Areas chosen were imitative ability, memory, and abilities to make gross shape, size, and brightness discriminations. In addition to being adaptable to the abilities of the children, these areas were thought to have special relevance to the purpose of the tests. Imitative abilities and memory are important in the education of these children, since imitative teaching techniques tend to be stressed in their training. Abilities to make simple discriminations in size, color, and shape are usually fundamental to the curriculum. Further, they are abilities which almost all children have had some opportunity to develop.

Age norms taken from the literature were studied to aid in meeting Criterion 2. The following procedures were used in the attempt to meet Criterion 3: Tests were not timed. Materials were three dimensional and relatively large. Surfaces were dull to minimize the reflection of light. Only minimal responses were required, such as simple choices among materials, which were placed

far enough apart so that a gross motion of hand, arm, or body sufficed to indicate the response.

In the search for interesting tests, (Criterion 4), toys and training materials chosen by the children in the classroom were noted. Sample materials were constructed, and were included in a group of toys presented to not less than twelve severely retarded children. Materials were rejected unless they were spontaneously selected by more than half of the children. Adequacy in meeting Criterion 5 could be estimated only on the basis of actual experimentation. Meanwhile, tests were designed to include a maximum of eight items.

Two methods were used to aid in meeting Criterion 6. First, and preferably, tests were chosen which seemed to contain within themselves the necessary experiences for performance. Failing this, tests were used if it could reasonably be assumed that materials and procedures were within the experience of all of the subjects.

To meet Criterion 7, verbal instructions to the subject were limited to simple words, accompanied by explanatory gestures in pantomime. No verbal responses were required, and the examiner manipulated the material at the child's direction, if necessary, (Criterion 8). An achromatic scale was used to meet Criterion 9, materials being white, black, and various grays in the Munsell Gray Scale. Materials were presented on a cloth of approximately the middle gray in the scale.

The Pilot Studies

A total of 115 mentally retarded subjects were used in the pilot studies.¹ Of these, 70 were between the ages of 5-0 and 12-6,

¹ Testing for the pilot studies was done by the principal investigator.

and were severely retarded according to the criteria used for the final sample, to be described later. Seven pilot studies were conducted with this group, each using ten subjects. In addition, eight subjects between the ages of 3-6 and 4-10, and six subjects aged 12-7 to 14-0 were tested, to aid in determining the most suitable age range for the final sample. Ten moderately retarded subjects, and eleven subjects whose retardation was medically attributed primarily to schizophrenia were tested on an experimental basis. The subjects were recruited from, and tested in, special schools for mentally retarded children.

A minimum of two tests in each area were included in the experimental battery, since it was expected that a number of the tests would prove to be unsuitable. The preliminary battery consisted of eleven tests.¹ These tests, and their evolution in the course of the pilot studies, are described in Appendix II. Procedures used in the collection of the final data were followed in the pilot studies; i.e., a standardized training period intervened between the first and second administrations of the tests, and alternate forms were used for the transfer situations. The later pilot studies were also devoted to experimenting with various time intervals for the retesting on which the retention scores were to be based.

¹ Five of the eleven tests in the experimental battery had been previously used in preliminary research in this area,⁽¹⁾ These were the Bang Test, the Egg Test, the Two-Buttons Test, the Light Test, and the Music Box Test. After a number of revisions, the last three were retained for the final battery, and two new tests, namely, the Buzz Test and the Three-Buttons Tests, were added, (see Appendices I and II).

Several transfer situations were tried for each test in the pilot studies. Since transfer was emphasized in this study, a word is in order concerning the concept as it was used here. Transfer was regarded as "the influence of prior learning upon the learning of, or response to, new related material."⁽²⁾ Transfer situations were constructed in the light of Thorndike's theory of "identical elements," and Judd's "generalization" theory. The former proposes that transfer depends on identical elements in the original learning and the new learning which it facilitates,⁽³⁾ and the latter hold that transfer occurs to the extent to which the individual generalizes his experience.⁽⁴⁾ Thus, transfer and original forms of the tests were designed to contain common elements, and, since ability to transfer tends to be limited in severely retarded children, only relatively few and comparatively simple changes were introduced. Transfer situations involving changes in shape, brightness, and size of materials, described below, are comparable in some respects to those used with both human and animal subjects, reported by Hebb,⁽⁵⁾ Harlow,⁽⁶⁾ Barnett and Cantor,⁽⁷⁾ and Houze, Zeaman, et al.⁽⁸⁾

For the transfer situations tried out in the pilot studies, different materials or procedures from those used in the original forms were used. In changes involving the materials, the following procedures were used: For tests of shape discrimination, brightness or size were changed, and shape held constant. In size discrimination tests, size was kept constant, and brightness or shape were changed. For tests of brightness discrimination, shape or size were changed, and brightness was held constant.

Various changes in procedures were also studied. In one method, the examiner repeated the original demonstrations, but omitted the gestures which originally accompanied them. In a second method, only

the original gestures were used, the demonstrations themselves being omitted. In a third method, the test itself was administered in a different place. In a fourth, the original materials were presented in reverse order.

Materials, procedures, and the order in which the tests were to be administered were determined in advance of each pilot study, and a list of criticisms was drawn up afterward. Tests were changed, substituted, accepted, or rejected according to their adequacy in meeting the test criteria. In some respects the behavior of the children was the best indicator of strengths and weaknesses in the tests. Inspection of the quantitative results was also used, and graphic analyses were added wherever possible.

The entire experimental battery was used in the first four pilot studies, pretraining, posttraining, and transfer scores being obtained for each test. The less suitable tests began to be dropped thereafter. The later pilot studies were devoted chiefly to improving the better tests, and to experimentation with the timing for the retention scores.

Results of the pilot studies suggested that above the age of 12-6 the tests tended to become too easy for the subjects to permit adequate discrimination among them, while below the age of 5-0 the tests appeared to be too difficult. Within that age range, however, the results of the pilot studies suggested that the effect of age on test performance was negligible. In order to remain within suitable age limits, the age range for the sample to whom the final battery was administered was set at 5-1 to 11-1.

The Subjects

The sample for the final battery was limited to children with available medical records, which included a diagnosis. Many of the

medical records came from special clinics for the mentally retarded, where the child was examined by a team. Some were supplied by physicians on the staff of the school. In a few cases, reports of the child's personal physician were used.

Medical diagnoses ranged from reasonably clear-cut statements of diagnostic category, such as Mongoloid, to highly general statements, such as "brain-injured." The latter was considered acceptable, in view of the purposes for which the diagnosis was obtained; namely, to limit the sample to children suffering from various pathological physical conditions known to produce mental retardation, and to permit a division into subgroups in order to analyze the influence of diagnostic category on test performance. For these purposes, statements such as "etiology unknown," or "unclassified" were not considered acceptable. Subjects were also excluded if retardation was medically ascribed primarily to "emotional problems."

The sample was also restricted to children who had received a psychological evaluation within the year. The evaluations included whatever standard intelligence tests could be used with the particular child, and were given by psychologists connected with the clinic or school. The Revised Stanford-Binet Scale was given to 60 of the subjects; the Merrill-Palmer Scale to 30; the Cattell Intelligence Scale for Infants and Young Children to 13; the Kuhlmann-Binet Scale to 3; and the Gesell Developmental Schedules to 2. Thirteen subjects could not, or did not, respond sufficiently to any of the tests to permit a meaningful quantitative score. The Vineland Social Maturity Scale, administered to the parent, was the only standard test that was used for all subjects.

No one standard intelligence test could be validly administered to all of the subjects. Most of the tests, (such as the Revised

Stanford-Binet and the Merrill-Palmer Scales), do not extend sufficiently far down in terms of mental age to cover the range of abilities investigated here. Further, most of the tests are not suitable for severely verbally and physically handicapped children. Tests which measure very early mental ages, (such as the Cattell Scale), are primarily tests of development at the lower end of the range, and also tend to be characterized by low reliability at these levels. In addition, these tests cannot be applied to the upper range of abilities studied here.

Since one measure of intellectual functioning for all of the subjects was necessary, both to determine the effect of this factor on test performance, and to delimit the sample in terms of intelligence, an estimated IQ was obtained for each child. This estimate was obtained from the psychologist of the school or clinic which the child attended, who was personally familiar with him, and who had evaluated him at least once within the year. The estimate was based on careful study, both quantitative and qualitative, of the child's psychological records. These records included results of whatever standard tests could be used, analyses and interpretations of the scores, observations of the child's behavior, and the psychologists' clinical evaluations of the child's abilities. In many cases, periodic re-examinations were included. That the psychologists based their estimated IQs largely on the scores of standard tests is shown by the fact that their estimated IQs correlated .92 with Stanford-Binet IQs were obtainable, and .81 with Vineland Social Quotients.

In limiting the sample to severely retarded subjects on the basis of psychological evaluations, the following criteria were used: An IQ below 50 on the Revised Stanford-Binet Intelligence

Scale was used in all cases where this scale could be meaningfully administered. Sixty of the subjects fell into this category. Failing this, an IQ under 50 on other standard intelligence tests was accepted. This criterion was used for 41 of the subjects. An estimated IQ below 50 was used for the 13 children who did not respond in a quantitatively meaningful way to any of the standard intelligence tests attempted.

Teachers' ratings were obtained from the child's current teacher, in whose class he had been for at least one month. A five-point rating scale was used, the teacher being asked to rate the child's learning ability on the basis of her experience with him in the classroom, in relation to the performance of the children who attended that school. The various populations of the different schools, all of which were special schools for mentally retarded children, were thought to be sufficiently similar to warrant the use of this frame of reference.

The final battery was administered to 114 severely mentally retarded children, living in the community, and attending special classes and/or clinics for the mentally retarded. Their estimated IQs ranged from 8 to 55,¹ with a mean of 36.18 and a standard deviation of 12.01. They had attended school from 0 to 36 months, mean 11.52 months, standard deviation 10.84. Their ages ranged from 61 to 143 months, with a mean of 95.16 months, and a standard deviation

¹ Estimated IQs tended to be several points higher than Stanford-Binet IQs. An estimated IQ up to 55 was therefore considered acceptable, provided the child's score on the Binet Scale was below 50.

of 23.48. There were 69 boys and 45 girls. All of them were medically stated to suffer from various pathological physical conditions known to produce mental retardation. Thirty-seven were Mongoloids. Fifty-five were cerebral-palsied or otherwise brain-injured. The remainder suffered from a variety of pathological conditions, such as cretinism, and cerebral "agenesis" or "hypogenesis."

The Final Battery

The five tests which comprised the final battery are briefly described below, in the order in which they were administered. Detailed descriptions, including photographs, procedures for administration, and copies of the scoring sheets, are given in Appendix I. As planned, the tests represented different levels of difficulty, in the sense that, for the group as a whole, higher mean scores were obtained on some tests than on others. Tests II and V were the most difficult, and Tests III and IV were the easiest, on the criterion of percent of items passed.

I. The Two-Buttons Test was primarily a test of imitative ability and visual and auditory memory. Materials consisted of a number of buttons, and two transparent boxes into which the buttons were to be dropped in alternating order. Any gross motion in the direction of the correct box or boxes into which the buttons should be placed constituted a correct response on the part of the subject. The examiner placed the buttons at the subject's direction, if necessary.

The test included six items, the buttons being presented in increasing numbers. Before each item, the examiner demonstrated the correct response, accompanying the drop of each button into the box with a verbal "Ping!" The subject was

then required to imitate the demonstration. In the transfer situation the actual demonstrations were omitted, all other procedures remaining the same.

This test was found to be a good beginning for the battery. A few preliminary clicks of the buttons in the boxes with the accompanying "Ping!" usually aroused immediate interest. Giving the child the buttons to play with in advance of testing proved to be an excellent way of establishing rapport.

II. The Buzz Test was chiefly a test of size discrimination. It consisted of round black discs of various sizes, attached to a gray board. When a disc of a comparable size was placed over its counterpart on the board, a continuous buzzing sound was produced. After pretest demonstrations, the subject was presented with the various discs, and asked "to make buzz" for each item. For a correct response the subject was required to indicate, by any unambiguous means at his disposal, the sample on the board which corresponded in size to the disc presented to him.

The test included six items, each consisting of the presentation of a single disc. In the transfer form, a series of white discs, corresponding in size to the black ones, were presented to the subject, to be matched to their black counterparts on the board.

This proved to be a satisfactory second test. The "buzz" was sufficiently different from the sounds in the previous test to provoke fresh interest and curiosity. Also, as one of the shorter tests, it was well suited to a period in which most of the subjects were still "settling down."

III. The Music Box Test was largely a brightness discrimination test. Round music boxes of equal size were used, one white, one black, one gray, one black and gray, and gray and white. The subject was shown in pretest demonstrations that only the gray box "made music." He was then asked "to find the music" at each presentation. Selection of the gray box, made in any unambiguous way, was the only response required.

The test consisted of six items. The gray box was placed in a different position for each item, and the other boxes were added in increasing numbers. Larger rectangular boxes, comparable in other respects to the round ones, were used for the transfer form.

This was the longest test in the battery. It was found to be most satisfactory when given in middle place, at a time when the children were usually sufficiently "settled" to be able to deal with a longer test, and when restlessness tended to be low.

IV. The Light Test, primarily a test of shape discrimination, consisted of a gray box with a plastic-covered window cut into each side. A single rectangular knob appeared on each side of the box, which, when pushed, lit a series of flashlight bulbs placed behind each window. The remaining knobs were non-functioning. The subject was asked, after pretest demonstrations, "to make light" on the various sides of the box. A correct response required selection of the rectangular knob for each item, made by the subject in any way possible to him.

The test consisted of eight items, each of the four sides of the box being presented twice. The knobs were used in various shapes and in increasing numbers. A second gray box,

with black knobs which corresponded in shape to the white ones, constituted the transfer form.

As one of the shorter tests, this was chosen to follow the longest one. It was placed toward the end of the battery because it was perhaps the best-liked of the tests, and was well suited to overcome developing restlessness and to restore flagging interest.

V. The Three-Buttons Test, a second test of brightness discrimination, involved sorting white, gray, and black buttons by dropping them into one of three transparent boxes. The examiner demonstrated the pattern to be followed by placing one sample button into each box. The subject was required to indicate the box containing the sample which corresponded to each of the buttons presented to him.

The test contained six items, the buttons being presented to the subject in increasing numbers. In the transfer form the position of the sample buttons was reversed, the subject now being required to place the buttons according to this pattern.

This test was placed at the end of the battery largely because a marked tendency toward perseveration resulted unless the two "buttons" tests were widely separated. Of the two, this one was found to be the better conclusion. It was short and interesting, making it particularly suitable as the final test.

The Test Administrations

The final battery was administered nine times.¹ Administrations A and B measured pre- and posttraining performance respectively, a standardized training period intervening between them. A different form of each test was used for Administration C, which yielded the transfer scores. Administrations A, B, and C were completed for each test before the next one was begun. The child was returned to his teacher or to his parent for the three-quarter of an hour interval between Administrations C and D.

Administration D measured retention without additional reinforcement. Pretest demonstrations and instructions were omitted. Training was not repeated, and the transfer forms were not included. This administration was given about three-quarters of an hour after completion of the original testing, since longer intervals of time for this method of retesting were found to result in scores which were too low for purposes of discrimination.

Administration E, which also measured retention, though under somewhat different conditions, was given one week after the original testing. Training periods and the use of the transfer forms were again omitted. However, pretest demonstrations and instructions were necessary, since it was found in the pilot studies that after this long an interval the majority of the children no longer remembered what was required of them.

Administrations F, G, H, and I were given approximately seven weeks after the original testing. They were repetitions of Administrations A, B, C, and D, respectively. Training was again included,

¹ All of the testing for which the final battery was used was done by qualified testers, who were given a manual for administering the tests were specially trained to administer it by the principal investigator, and were supervised by her in actual administrations with at least twelve subjects each. Scores obtained by these subjects were not included in the results.

and both forms of the tests were used.

General Procedures

General procedures described here include those used for screening the subjects, for testing, for scoring, and for training.

Procedures for Screening

The subject's sight and hearing were evaluated by the following procedures: Medical records, teachers' reports, psychological evaluations, and questionnaires supplied by parents were studied. Trained personnel familiar with the child were consulted. If these sources gave no indications that the child's vision and hearing were seriously impaired, the examiner undertook special screening procedures for the specific purposes of the tests.

Before the child was brought into the room, a number of the smaller materials were placed on the table, covering approximately the area to be used in testing. If the child picked up these materials, reached toward them, or glanced from one to another, it was assumed that his vision was adequate for test purposes.

The child's responses to remarks made by the examiner from various positions behind the child were also noted. Where sound was involved in the test materials themselves, these materials were first presented in several positions outside the child's range of vision. If he turned, reached, or glanced in these directions, it was assumed that his hearing was adequate for purposes of the tests.

Procedures for Testing

Subjects were tested individually, in the school or clinic which they attended. The number of sessions used to complete the administration of the battery depended on the needs of the particular

child. In most cases, the first three administrations could be completed in one session.

The child's handedness was determined in advance, if possible. In the few cases where this was not possible, right handedness was assumed. The test items were presented toward the child's favored hand, the middle point of the item corresponding roughly to the midline of the shoulder.

If the child could not, or did not, manipulate the materials himself, the examiner did so at the child's indication. The subjects were thus given equal opportunities to see their completed responses, and to make or indicate spontaneous corrections. The examiner also corrected, or correctly completed, all responses, so that the subjects might see the correct response the same number of times.

Procedures for Training

The subject was trained to the correct response only. Training to the avoidance of incorrect responses was not included. Extremely simple training procedures were used, in order that the majority of the children would achieve some successes in this phase of the testing. Training consisted of demonstrations of the correct response by the examiner, imitation of this demonstration by the subject with cues supplied by the examiner, and prevention or immediate correction of errors. Again, the examiner manipulated the materials at the child's direction, if necessary. Most of the subjects were thus able to give a number of correct responses during the training periods. The subject was trained to a maximum of three times on each item in the test. The special training procedures which were followed for the various tests are described in Appendix I.

Procedures for Scoring

A separate sheet was used in scoring each test, separate scores being obtained for each of its administrations. Numbers and letters were used on the scoring sheets, to show the materials each item contained, and the order in which they were to be presented. The subject was given one point for each correct response. Copies of the scoring sheets are included in Appendix I.

Summary of Chapter

Criteria were set up to aid in the construction and development of the tests. These required that tests be within the abilities of the subjects and suited to their verbal, motor, and sensory handicaps; arouse and maintain their interest; and be adequate for purposes of discrimination. Areas chosen for study were imitative abilities, memory, and shape, brightness, and size discrimination. An experimental battery, consisting of eleven tests, was constructed, and administered in a series of pilot studies to 115 mentally retarded children. The chief aim of the pilot studies was to bring the tests increasingly into accord with the criteria.

The sample to which the final battery was administered consisted of 114 severely mentally retarded children, living in the community, and attending special classes and/or clinics for the mentally retarded. Their estimated IQs ranged from 8 to 55, mean 36.18, standard deviation 12.01. Their ages ranged from 61 to 143 months, mean 95.16 months and standard deviation 23.48. They had attended school from 0 to 36 months, mean 11.52 months, standard deviation 10.48. There were 69 boys and 45 girls. All of them were medically stated to suffer from various pathological conditions known to produce mental retardation.

The final battery consisted of five tests, each of which was administered nine times. A pretraining, a posttraining, a transfer, and various retention scores were obtained for each test. The subjects were tested individually. General procedures for selecting and screening the subjects, and for testing, training, and scoring are described.

References

1. Schucman, H., "A Method for Measuring Educability in Severely Mentally Retarded Children: A Preliminary Study," The Training School Bulletin, Vol. 54, Nos. 3 and 4, 1957, and No. 1, 1958.
2. Craft, L. W., Schneirla, T. C., Robinson, E. E. and Gilbert, R.W., Recent Experiments in Psychology, New York: McGraw-Hill Book Co., 1950, P. 296.
3. Hilgard, E. R., Theories of Learning, New York: Appleton-Century-Crofts, 1948, P. 431.
4. Webb, L. W., "Transfer of Training," in C. E. Skinner (ed.), Educational Psychology, New York: Prentice-Hall, 1950, P. 266.
5. Hebb, D. O., The Organization of Behavior: A Neuropsychological Theory, New York: John Wiley and Sons, 1949, P. 14.
6. Harlow, H. F., "The Formation of Learning Sets," Psychol. Review, Vol. 56, 1949, p. 51.
7. Barnett, C. D. and Cantor, G. N., "Discrimination Set in Defectives," Amer. J. Ment. Defic., Vol. 62, 1957, p. 334.
8. House, B., Zeaman, D., Orlando, R. and Fischer, W., Learning and Transfer in Mental Defectives, Univ. of Conn., 1957.

III

RESULTS

The experimental data consisted of scores obtained from nine administrations of a battery of five tests. Administration A measured initial or pretraining performance. Administrations B, C, D, and E measured different aspects of learning. Administrations B and C yielded posttraining and transfer scores respectively. Administrations D and E produced retention scores, at different times and under different conditions. Administrations F, G, H, and I were repetitions of Administrations A, B, C, and D, given at a later date. The sample consisted of 114 subjects, all of whom received the first four administrations. Of the original subjects, 102 were still available for the fifth administration, and 36 for the last four.

This chapter presents the results.¹ It deals with test-retest comparisons, inter- and intratest relationships of the battery, comparisons of scores obtained by the various administrations, the effects of population variables on test performance, and the relationships of the scores with external criteria.

¹ Complete statistical tables are on deposit at the Gould Memorial Library, New York University, University Heights, New York 53, New York. Refer to Supplement to Measurement of the Educability of Severely Mentally Retarded Children, 1959. The supplement is available on interlibrary loan.

Test-Retest Comparisons

Thirty-six of the original subjects, all of whom had received the first five administrations of the battery, were still available for retesting after an interval of seven weeks.¹ Product-moment correlations were computed between their scores on comparable administrations of each test, in order to determine the test-retest stability of the measures. These coefficients, which were all significant at the .001 level of confidence are shown in Table 1. It can be seen that those yielded by pretraining performance, (Administrations A and F), were always lower than those obtained

Table 1
Product-moment Correlation Coefficients*
Obtained by Test-Retest Comparisons
(N = 36)

Administrations	Test I	Test II	Test III	Test IV	Test V
A and F	.66	.82	.85	.79	.80
B and G	.93	.95	.94	.92	.95
C and H	.93	.90	.91	.92	.96
D and I	.95	.90	.95	.92	.95

* $p < .001$

1

To prevent possible contamination, the original scoring booklets were taken from the testers after completion of Administration E. Also, the testers were not informed in advance of the later testing that any further testing was anticipated. New scoring booklets were used for Administrations F, G, H, and I. In no case did the testers see the original scores again until all testing was completed.

between comparable learning scores. The former ranged from .66 to .85 and the latter from .90 to .96, with no overlap between the ranges.¹

Intertest and Intratest Relationships

To determine the internal consistency of the battery, the scores were analyzed for intertest and intratest relationships.

Intertest Relationships

Product-moment correlations were computed between the same administrations of the different tests to determine intertest relationships; i.e., scores on Administration A of each test were compared with scores of Administration A of every other test, and so on. Ten coefficients were thus obtained for Administrations A, B, C, D, and E respectively.

All of the coefficients were significant at the .001 level of confidence. The ten coefficients yielded by initial test performance, which ranged from .31 to .63, were lower than those produced by the forty comparisons based on learning scores, which ranged from .72 to .87, with no overlap between the ranges. The likelihood of this distribution occurring by chance is of the order of $p < .001$.

¹ To test the null hypothesis that the pretraining correlations were no different on the average from those obtained between the learning scores, one can evaluate the likelihood that the A-F correlations would be consistently lower than those between the learning scores in five independent trials. This is of the order of $(1/4)^5$, or $1/1024$. The null hypothesis can therefore be rejected at the .001 level of confidence.

Intratest Relationships

Product-moment correlations were also computed between different administrations within the same tests; i.e., between Administration A, B, C, D, and E of the same test. Ten coefficients were thus obtained for each test. All of them were significant at the .001 confidence level. The twenty coefficients yielded between pre-training performance and the various learning scores within each test were lower than those obtained by the thirty comparisons based on learning scores. The former ranged from .50 to .89 and the latter from .86 to .97, overlap between the ranges occurring in only one instance, ($p < .001$).

Comparisons of Scores Obtained by the Various Administrations

To determine whether the various administrations of the battery had produced significantly different results, a repeated measurements analysis of variance was performed for each test, between scores obtained by pretraining, posttraining, transfer, and retention performance, as obtained on Administrations A, B, C, and D respectively.¹ All of the subjects were used for these comparisons. The results demonstrated that highly significant differences were obtained by the various administrations of every test, the F_s ranging from 94.90 to 163.97, ($p < .001$). With one exception,

¹ Scores from Administration E, the second measure of retention, were omitted here. Scores on Administration E, which was not given to all of the subjects, were very similar to those produced by Administration D, significant differences between the two means being found only on Test I, ($p < .05$). Nonsignificant mean differences were found in all other instances. All differences between initial test scores and those on Administration E, like those between initial performance and Administration D, were significant at the .001 level.

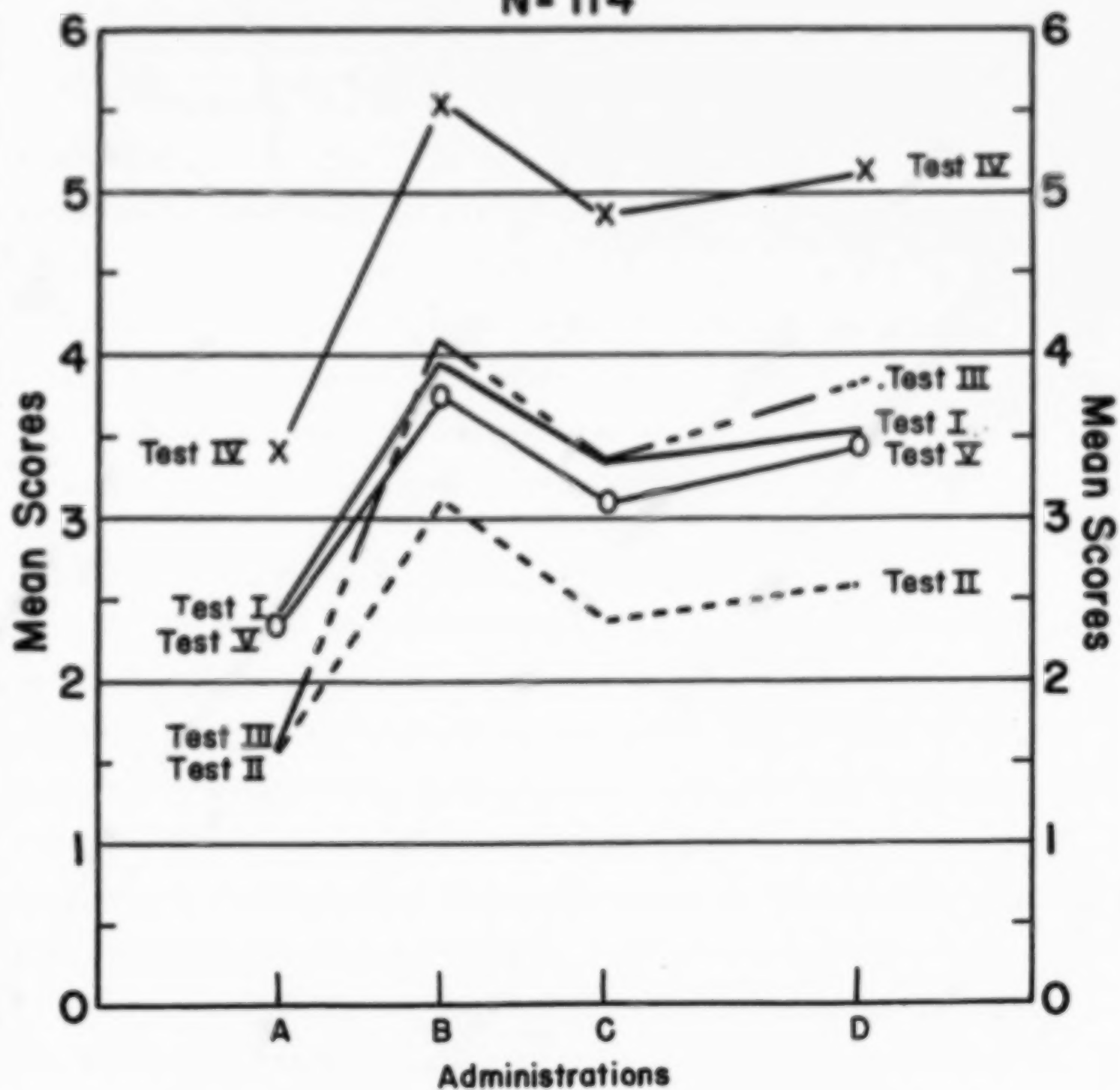
(between posttraining and retention scores on Test III), t_s between the means showed that significant differences were obtained between all administrations of all tests, the t_s ranging from 2.09, ($p < .05$) to 20.35, ($p < .001$).

The mean scores obtained on the four administrations of each test are shown in Figure 1. It can be seen that the curves for every test followed a highly consistent pattern, and one in essential accord with the various types of behavior which the administrations were designed to measure. The largest differences were always found between pretraining performance on the one hand and learning performance on the other. Of these differences, the greatest were always between pre- and posttraining scores. Transfer scores were always significantly lower than posttraining scores, ($p < .001$). Retention scores were always significantly higher than transfer performance, ($p < .001$), and with one exception significantly lower than posttraining scores.

A second series of repeated measurements analyses of variance were performed between the scores obtained by the 36 subjects who received all administrations of the battery, including Administrations F, G, H, and I. The F_s between administrations ranged from 12.16 to 29.44, ($p < .001$), showing that highly significant differences were obtained between these administrations. The t_s between the means of Administration F and the previous administrations showed the following: Except on Test I, the mean scores on Administration F were higher than those on Administration A, with t_s ranging from 3.83 to 6.06, ($p < .001$). The mean scores of Administration F were always significantly lower than those of Administration B, with t_s from 2.83, ($p < .01$) to 5.81, ($p < .001$). Mean scores of Administration F were significantly lower than those of Administration D in three out of five instances.

Differences in Mean Scores for each Administration of each Test

Figure 1
N = 114



A = Pretraining

C = Transfer

B = Posttraining

D = Retention

The mean scores obtained by the 36 subjects on Administrations A through I of each of the tests are shown in Figure 2. It can be seen that highly consistent learning curves were obtained across time by Administrations A, B, C, and D on the one hand, and Administrations F, G, H, and I on the other.

Summary

In comparing the scores of all of the subjects on pretraining, posttraining, transfer, and retention performance, highly significant differences were found on all of the tests. The learning curves followed a highly consistent pattern, and one which would be logically anticipated from the different types of behavior the various administrations were designed to measure. A similarly consistent and logical pattern was obtained across time in the learning curves of the 36 subjects who had received all administrations of the battery.

The Effect of Population Variables on Test Performance

The effect on test performance of five independent variables was analyzed; sex, time spent in school, age, diagnostic category, and intelligence (as measured by estimated IQs). Their influence was analyzed on pretraining, posttraining, transfer, and retention performance of all of the subjects, as measured by Administrations A, B, C, and D respectively. Scores obtained on Administration E were also included in later analyses.

Before analyzing the influence of the independent variables on the scores, multiple correlations were computed, to determine the extent to which these factors together accounted for the variation in the test scores. The coefficients are shown in Table 2.

Mean Scores for Various Administrations

Figure 2

N=36

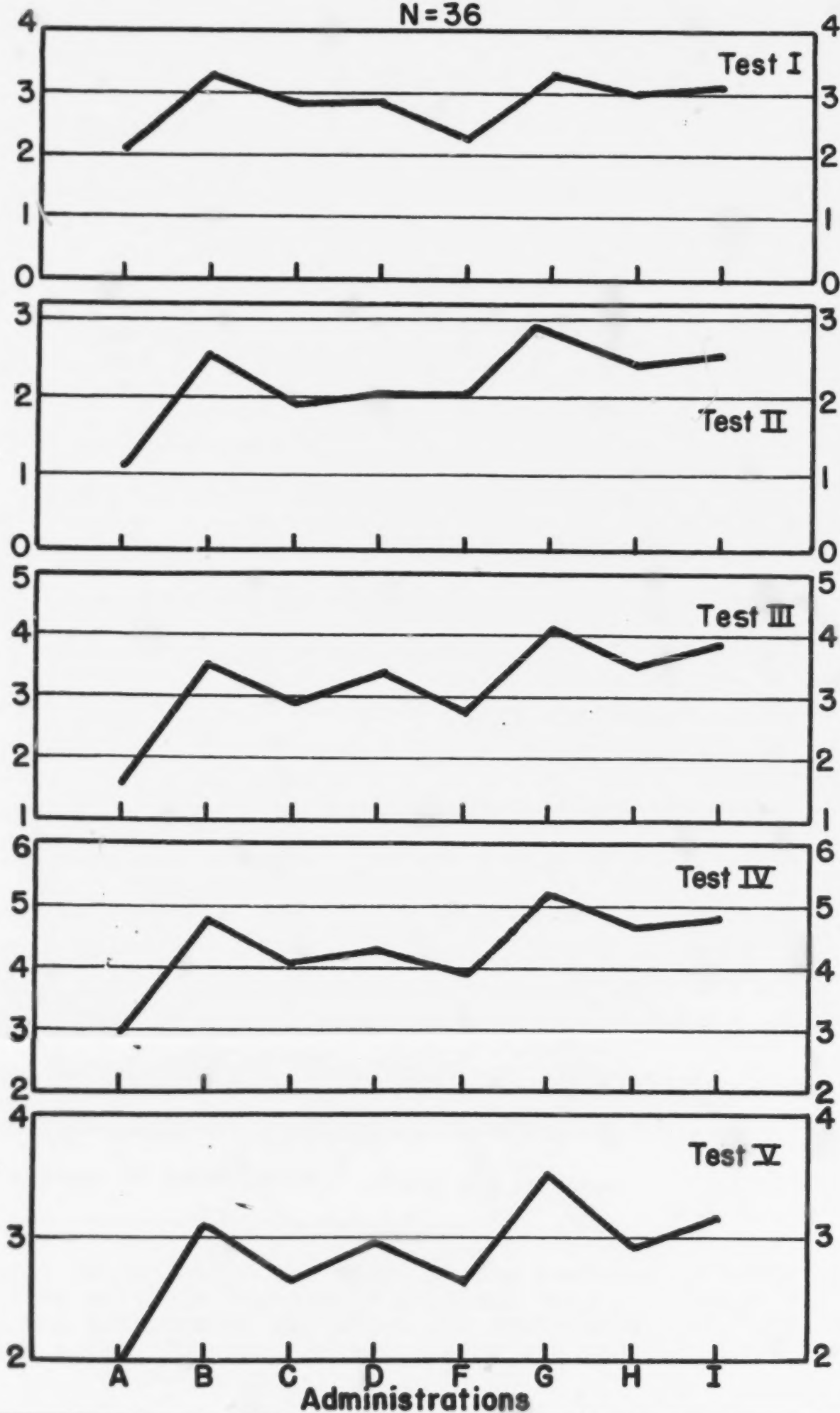


Table 2
Multiple Correlation Coefficients*
Between the Five Independent Variables and Administrations
of Each Test

(N = 114)

Administration	Test I	Test II	Test III	Test IV	Test V
A	.72	.72	.48	.67	.78
B	.88	.87	.86	.86	.87
C	.90	.85	.86	.85	.85
D	.83	.84	.84	.86	.86

* p < .001 in all cases

It can be seen that the coefficients between pretraining scores and the independent variables, which ranged from .48 to .78, were always lower than those produced between the learning scores and the independent variables, which ranged from .83 to .90, with no overlap between the ranges, (p < .001). Expressed as coefficients of determination, (R^2), these results indicate that the five independent variables together accounted for 23% to 61% of the variation in initial test performance, and 69% to 80% of the variation in learning performance.

The effect of the variables of sex, time spent in school, age, diagnostic category, and intelligence on Administrations A, B, C, and D of each test was analyzed by a series of complex factorial design analyses of covariance,¹ using all of the subjects. The

¹ The following design was used for the analyses of covariance: Twenty-five multiple regression analyses were performed, one for each of the independent variables and Administrations A, B, C, and D of each test. In each series, one of the independent variables

analyses were repeated for each independent variable and each test, using the actual scores, in a series of mixed two-factor analyses of variance,¹ in which the scores of Administration E were included.

The Effect of Sex on Test Performance

The analyses of covariance, based on 69 boys and 45 girls, demonstrated that the factor of sex, both alone and in interaction with the test administrations, did not significantly influence performance on any of the tests. A nonsignificant effect of this variable was also shown by the analyses of variance, which were based on 24 boys and 24 girls.

(Continued)

was used as a classification factor, the other four being treated as predictors. The range of actual values was used for the variables of age, time in school, and intelligence when used as predictors. For sex, boys were assigned a score of 1 and girls a score of 2. For diagnostic category, brain-injured subjects were assigned to level 1, a miscellaneous group to level 2, and Mongoloids to level 3. (The rationale for the levels used for the latter variable is stated in the section devoted to its effect on the test scores). A predicted score was thus obtained for each subject on each administration of each test. This score was then subtracted from the subject's actual score, producing a series of 114 residuals, one series for each administration of every test. Using the residuals thus obtained, a series of analyses of variance were performed between levels of each of the independent variables and the four administrations of each test. The particular series of residuals used for each of these analyses were based on the covariance of four of the independent variables, excluding the one now being used as the classification factor.

The residuals were computed on an IBM 704 Digital Computer, and the analyses of variance based on the residuals were computed on an IBM 650 Digital Computer.

¹ In the analyses of variance, groups were proportionalized so that nonsignificant differences were obtained for pertinent population variables other than the one under consideration, t-tests and chi-squares being used to test the significance of the differences. Although this procedure had the disadvantage of reducing the Ns, a reduction which was also necessitated by the inclusion of Administration E, which was not given to all of the subjects, the use of this technique was thought to be warranted in view of the additional information obtained.

The Effect of Time in School on Test Performance

Three levels were used for the variable of time spent in school. The groups were based on length of time the subject had actually attended school. Levels used in the analyses of covariance are shown in Table 3. Subjects with little or no school were assigned to the lowest level. The middle level consisted of subjects

Table 3
Levels of Time in School
Used in Analyses of Covariance

Level	Range in Months	Mean in Months	Number of Subjects
1	0 - 2	.97	33
2	3 - 13	7.48	40
3	14 - 38	23.95	41

with moderate amounts of time in school. The highest level included subjects who had attended school for relatively long periods. The analyses of covariance demonstrated that the factor of time spent in school did not significantly affect the scores of any of the tests, either alone or in interaction with administrations.

The levels of time in school used in the analyses of variance are shown in Table 4. These analyses also showed that this variable did not significantly influence performance on any of the tests, either as a main effect or in interaction with administrations.

Table 4
Levels of Time in School
Used in Analyses of Variance

Level	Range in Months	Mean in Months	Number of Subjects
1	0 - 2	1.04	28
2	3 - 13	7.58	26
3	14 - 33	21.53	21

The Effect of Age on Test Performance

The three levels used for the factor of age corresponded to three general levels of school readiness for this population. The groups were older than those for "normal" children at corresponding levels, which is usual with severely retarded subjects. The youngest level corresponded to a preschool group. The middle level constituted a beginning and early school group. The oldest group was ready for more sustained and concentrated training on the basis of age. The levels of age used in the analyses of covariance are shown in Table 5.

Table 5
Levels of Age
Used in Analyses of Covariance

Level	Range in Months	Mean in Months	Number of Subjects
1	61 - 83	71.24	42
2	84 - 112	96.00	42
3	114 - 143	127.47	30

Results of the analyses of covariance showed that age did not significantly affect scores on Tests I or IV. No significant interaction was found between age and administrations on any test. However, significant F s were obtained between age levels on Test II, (6.42, $p < .01$), Test III, (3.63, $p < .05$), and Test V, (3.42, $p < .05$), showing that age significantly influenced scores on these tests as a whole but not on specific administrations. The t s between the mean residuals of the age levels on Tests II, III, and V, which ranged from 2.28 ($p < .05$) to 4.40, ($p < .001$), demonstrated that the youngest group was significantly inferior to the middle level on Tests II and III, and to the oldest level on all three tests. No significant differences were obtained between the performance of the middle and oldest groups, or between any of the age levels on specific administrations.

Essentially the same results were yielded by the analyses of variance. The age levels used in these analyses are shown in Table 6. As before, no significant effect of age was found on scores of

Table 6
Levels of Age
Used in Analyses of Variance

Level	Range in Months	Mean in Months	Number of Subjects
1	61 - 83	71.00	37
2	84 - 112	96.39	36
3	115 - 143	126.81	21

Tests I or IV, and no significant interaction effect was obtained on any test. However, in this case the F between levels on Test V failed

to reach significance. As in the analyses of covariance, the F s between levels on Tests II and III were significant, (4.37 and 4.51 respectively, $p < .05$). Between-level t s on Tests II and III, ranging from 2.20, ($p < .05$) to 2.74, ($p < .01$), showed that the mean score of the youngest group was significantly lower than that of the middle or oldest levels. As in the analyses of covariance, no significant differences were found between the means of the middle and oldest groups, or between the levels on specific administrations.

In summary, the youngest subjects were inferior in performance on some of the tests as a whole. No significant differences were found between the scores of the middle and oldest groups.

The Effect of Diagnostic Category on Test Performance

Levels used in the analyses of the effect of diagnostic category on test performance were based on the extent of handicap and impairment of general level of functioning.¹ Brain-injured² subjects, who were assigned to the lowest level, were considered to be most handicapped in that various types of physical handicaps were present in addition to mental retardation. Mongoloids were assigned to the highest level. They were thought to be the least handicapped since, as a group, their motor coordination was comparatively good and their

¹ Levels used for the variable of diagnostic category and assignment of subjects to them were approved by James Toolan, M.D., of the Department of Psychiatry, New York University College of Medicine and Bellevue Hospital.

² "Brain-injured" subjects were subjects who had suffered trauma to the brain. Cerebral-palsied children were placed in this group, but were not treated separately because they do not constitute a unitary classification, and the diagnoses did not always include elaboration. Further, some of the subjects medically stated to be brain-injured, in general terms, may have been cerebral-palsied. Despite the disadvantages of using a general "brain-injured" category, no rationale for meaning-subdivision within the group was found.

behavior relatively stable. The middle level consisted of subjects with various medical diagnoses, such as cretinism, or cerebral "agenesis" or "hypogenesis," in insufficient numbers to be treated as separate groups, and without medical evidence to warrant the assumption of the superimposition of brain-injury.

The analyses of covariance were based on 55 subjects in level 1, 22 in level 2, and 37 in level 3. Results demonstrated that the factor of diagnostic category did not significantly influence scores on Tests II, III, or V. No interaction effect was obtained on any of the tests. However, significant F s between the levels were obtained on Test I, (4.54, $p < .05$) and Test IV, (4.89, $p < .01$), demonstrating that this variable significantly affected performance on these tests as a whole. The t s between mean residuals of the levels showed that the Mongoloids performed significantly better than the brain-injured subjects on Tests I and IV, (2.99, $p < .01$, and 2.24, $p < .05$, respectively). On Test IV, the Mongoloids also performed significantly better than the miscellaneous group, (t , 2.96, $p < .01$). No significant differences between levels were found on specific administrations.

The analyses of variance, which were based on 43 subjects in level 1, 21 in level 2, and 17 in level 3, yielded essentially similar results. Again, no effect of diagnostic category was found on Tests II, III, or V, and no significant interaction effects were obtained. Here, however, the between-level F on Test I did not reach significance. As in the analysis of covariance, a significant F between levels, (3.46, $p < .05$), was obtained on Test IV, the t s between levels, (2.03 and 2.52, $p < .05$), showing that the mean score of the Mongoloids was significantly higher than that of the other groups.

Again, no significant differences between the levels were obtained within specific administrations.

In summary, the factor of diagnostic category did not significantly affect performance on most of the tests. However, the Mongoloids were superior to the brain-injured subjects on some of the tests, and in one case were also superior to a miscellaneous group.

The Effect of Intelligence on Test Performance

The division of the subjects into levels of intelligence was based on the division of the range of estimated IQs approximately into thirds, creating a low, a middle, and a high group. Levels used in the analyses of covariance are shown in Table 7.

Table 7
Levels of Intelligence
Used in Analyses of Covariance

Level	Range of Estimated IQs	Mean of Estimated IQs	Number of Subjects
1	8 - 25	16.54	24
2	27 - 39	34.25	36
3	40 - 55	46.20	54

Analysis of covariance clearly demonstrated that the factor of intelligence was highly significant in influencing performance on every test, both alone and in interaction with the administrations.

The F_s between the levels of intelligence ranged from 48.64 to 68.26, ($p < .001$). The t_s applied to the mean residuals of the levels

ranged from 3.61 to 10.94, ($p < .001$). In every case, the means for the high group were significantly higher than those of the low or middle levels, while the means of the low level were significantly lower than those of the middle group. These results demonstrate the highly significant effect exerted by intelligence on every test as a whole.

The t_s between levels on specific administrations demonstrated that the high subjects always performed significantly better than the low group in learning performance, with t_s ranging from 5.05 to 6.48, ($p < .001$). With two exceptions, the high group was also significantly superior in learning performance to the middle group, the t_s ranging from 2.36, ($p < .05$) to 4.72, ($p < .001$). With three exceptions, the low group was significantly inferior in learning performance to the middle group, with t_s from 2.25, ($p < .05$) to 3.73, ($p < .001$). In brief, significant differences between the means of the intelligence levels were obtained in 40 of the 45 comparisons based on learning performance. On the other hand, the t_s between the levels on initial test performance, which were always lower than those for learning performance at corresponding levels, failed to reach significance in 8 of the 15 comparisons.

Of particular interest is the significant interaction between the levels of intelligence and the administrations, which was found on every test. The interaction F_s ranged from 3.44, ($p < .01$) to 10.97, ($p < .001$), showing that the factor of intelligence had exerted a significant differential influence on the different test administrations.

The interaction t_s demonstrated that, when the factors of age, sex, time in school, and diagnostic category were covaried, learning

performance was significantly more closely related to the level of intelligence than was initial test performance on every test. The ts ranged from 2.29, ($p < .05$) to 6.31, ($p < .001$). With one exception, nonsignificant differences were obtained between the various learning scores.¹

In summary, analysis of covariance clearly demonstrated the highly significant influence of intelligence on the scores of every test, both as a main effect and in interaction with the administrations. The different levels of intelligence were always more sharply differentiated by learning than by pretraining performance.

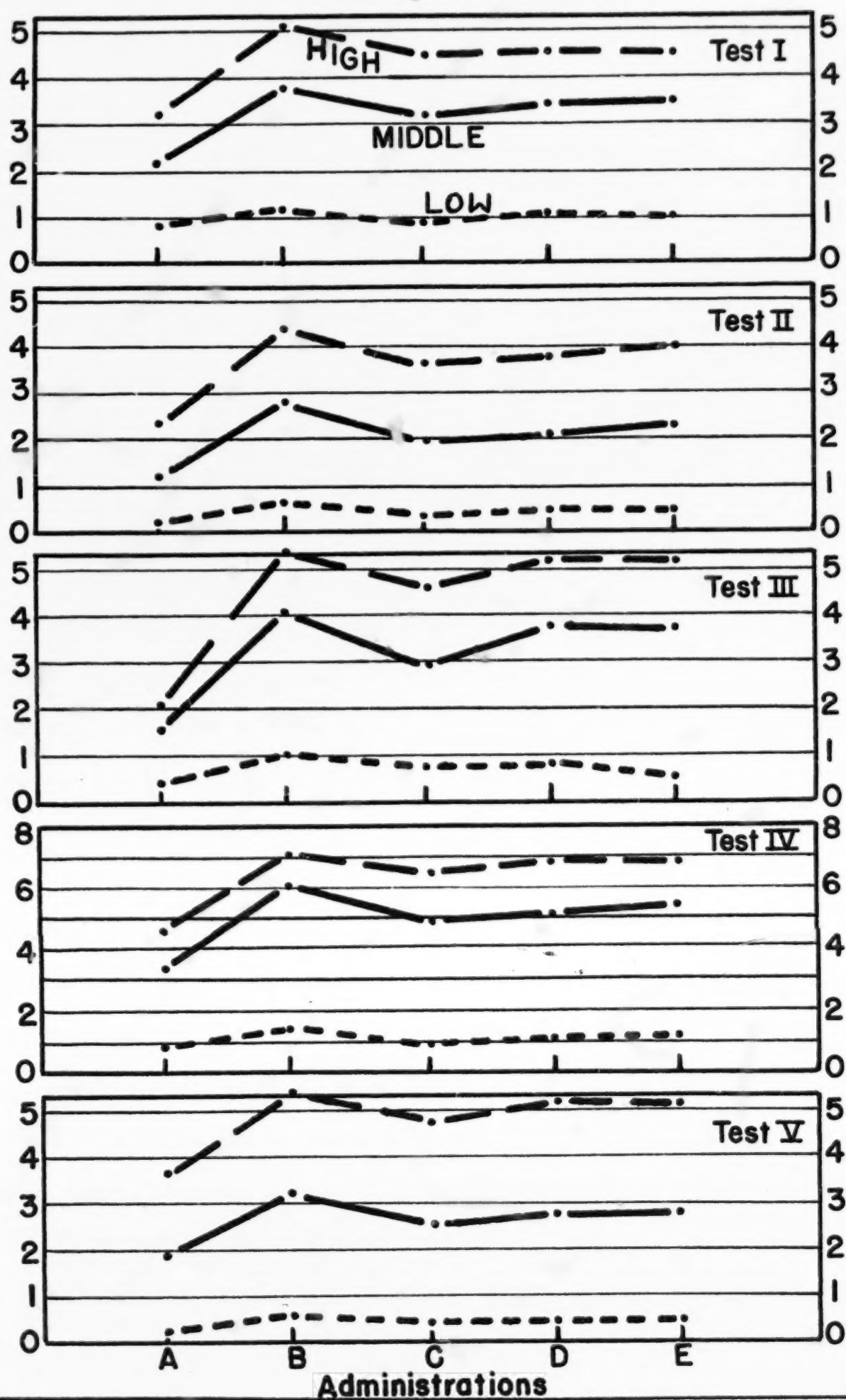
Results of the analyses of variance also showed a highly significant effect of intelligence on performance on every test, both alone and in interaction with the administrations. These results are clearly shown in Figure 3, which presents the mean scores of the different intelligence levels on the various test administrations. The levels used in these analyses are shown in Table 8.

As in the analyses of covariance, the Fs between the different levels were highly significant on every test, ranging from 68.62 to 121.92, ($p < .001$). Again, the ts between levels were all significant at the .001 level, ranging from 4.53 to 15.52. As before the high group always performed significantly better than the low or middle levels, while the low group was always significantly inferior to the middle level on every test as a whole. Also as before, ts

¹ The nonsignificant differences between these learning scores demonstrate only approximately equal accuracy of prediction, since the analyses of covariance were based on the differences between actual and predicted scores. The accuracy of the prediction does not reflect the shape of the learning curves based on the actual scores obtained by the different levels of intelligence. These are shown in Figure 3.

Mean Scores of Analyses of Variance for Levels of Intelligence

Figure 3



demonstrated that the mean scores of the high group were always significantly greater than those of the low group on all learning scores, with ts ranging from 5.15 to 7.76, ($p < .001$), while those

Table 8
Levels of Intelligence
Used in Analyses of Variance

Level	Range of Estimated IQs	Mean of Estimated IQs	Number of Subjects
1	8 - 25	15.95	21
2	27 - 39	34.10	31
3	40 - 55	46.40	50

of the low group were always significantly lower than those of the middle level, ts ranging from 2.45, ($p < .05$) to 5.60, ($p < .001$). With one exception, the middle group was significantly inferior to the high group in learning performance, with ts ranging from 2.22, ($p < .05$) to 4.47, ($p < .001$).

Thus, significant differences between the mean scores of the various levels of intelligence were found in 59 of the 60 comparisons based on learning performance. The ts obtained on initial test scores were always lower than those based on the learning scores, failing to reach significance in 4 of the 15 comparisons.

As in the analyses of covariance, a highly significant interaction between intelligence level and administrations was found on every test. The interaction Fs ranged from 7.09 to 20.95, ($p < .001$). Here, the differences in actual scores obtained by the different

intelligence levels on the various administrations were clearly shown by the within-level ts, which showed the following:

For the low group, the effect of training was minimal. The only significant learning gains were obtained on posttraining performance, where the mean scores were significantly higher than those of initial performance on four of the five tests, ts ranging from 2.06, ($p < .05$) to 2.62, ($p < .01$). The low group did not transfer the training or retain the learning significantly on any test.

For the middle group, all mean learning scores were significantly higher than those of initial test performance, the ts ranging from 4.14 to 13.48, ($p < .001$). It can therefore be said that the middle group learned in response to training on all tests, and were also able to transfer and retain the learning to a significant extent in all cases. Within the various learning scores, significant differences were found in 17 of the 30 comparisons, with ts ranging from 2.07, ($p < .05$) to 6.14, ($p < .001$).

For the high group, highly significant differences were always obtained between mean scores of pretraining and learning performance, the ts ranging from 10.83 to 51.35, ($p < .001$). Thus, the high group learned, transferred the training, and retained the learning to a highly significant extent on every test. The magnitude of the scores of the high group was always greater than that of the middle group. Within the various learning scores, significant differences were found in 21 of the 30 comparisons, with ts ranging from 2.20, ($p < .05$) to 18.92, ($p < .001$).

In summary, analysis of variance demonstrated a highly significant influence of intelligence on test performance, both alone and in interaction with administrations. The low group achieved minimal

learning gains, and obtained no significant transfer or retention scores. The middle and high groups obtained significant learning, transfer, and retention scores on all tests, the magnitude of the mean scores of the high group being always greater than those of the middle level. On every test, intelligence was found to be more closely related to learning performance than to initial test performance.

Still another demonstration of the greater influence of intelligence on learning performance as opposed to pretraining performance can be found in the coefficients of determination based on product-moment correlations between this factor and the four administrations of each test.¹ While the variables of sex, time in school, age, diagnostic category, and intelligence together accounted for 23% to 61% of the variation in initial test performance, (see Table 2), intelligence determined from 18% to 55%. On learning performance, the five factors together accounted for 69% to 80% of the variation, 62% to 76% being attributable to intelligence. These results are further evidence of the close relationship existing between intelligence and learning performance.

Summary of Effects of Population Variables on the Scores

The variables of sex and time in school did not significantly affect performance on any test. Age and diagnostic category had a limited effect on some of the tests. Intelligence was the only factor which was significantly related to performance on all of the

¹ The product-moment correlations on which the coefficients of determination were based were performed by an IBM 704 Digital Computer, as a by-product of the analysis of covariance.

tests, both as a main effect and in interaction with administrations. Learning performance was always a more sensitive measure of differences in intelligence level than was initial performance. Among the learning scores, transfer and retention performance were the most sensitive to differences in intelligence level.

Comparisons of Test Scores and External Criteria

Before comparing the scores with teachers' ratings, psychologists' estimated IQs,¹ and Vineland Social Quotients, product-moment correlations were computed to determine the relationships of these criteria to each other, and to IQ obtained on the Revised Stanford-Binet Intelligence Scale, where obtainable. The coefficients, which were all significant at the .001 level, (see Table 9), show that the criteria were not unrelated. Estimated IQs demonstrated

Table 9
Product-Moment Correlation Coefficients
Between Estimated IQs, Stanford-Binet IQs, Teachers'
Ratings and Vineland SQs

	Teachers' Ratings	Estimated IQs	Vineland SQs
Estimated IQs	.82***	-	-
Vineland SQs	.72***	.81****	-
Binet IQs	.54*	.92**	.61**

p for all $r_s < .001$

* $N = 51$

** $N = 60$

*** $N = 99$

**** $N = 114$

¹ The close relationship between estimated IQs and test performance has already been demonstrated. They were included here because of the added information supplied in this context. The correlations were computed on an IBM 650 Digital Computer, the inter- and intra-test correlations reported earlier being computed at the same time.

the greatest commonality with the other measures. The closest relationship was found between estimated IQs and Binet IQs, (.92). This relationship was to be expected, since the psychologists tended to base their estimates chiefly on Binet IQs. The lowest correlations were obtained between teachers' ratings and Binet IQs.

In order to determine the reliability of estimated IQs and of teachers' ratings, two such measures were obtained for the same subjects wherever possible. In no case did one rater know the estimate given by the other until both were secured. Eleven subjects had been evaluated twice by different psychologists within the year, one at a clinic and one at a school. Product-moment correlations computed between these estimated IQs yielded a coefficient of .87, ($p < .001$). In 25 cases, two teachers had taught the same child for at least a month during the past year. Product-moment correlations between their ratings produced a coefficient of .85, ($p < .001$).

Scores from Administrations A, B, C, D, and E of each test were correlated with teachers' ratings, estimated IQs, and Vineland Social Quotients. The coefficients are shown in Table 10. Although all of the coefficients were significant, ranging from .31 to .87, it can be seen that for each test, the coefficients between initial test scores and the criteria were always lower than those between learning performance and the criteria. In order to determine the significance of these differences, t s were applied between the coefficients obtained for pretraining performance and the criteria on the one hand, and learning performance and the criteria on the other.

In all cases, learning performance was found to be more significantly related to estimated IQs than was pretraining performance,

Table 10

Product - Moment Correlation Coefficients between Test Scores and
Estimated IQs, Teachers' Ratings, and Vineland Social Quotients

Test I (N=102) (N=99) (N=102)			Test II (N=102) (N=99) (N=102)				
Admin.	Esti- mated IQs	Teachers' Ratings	Vine- land SQs	Admin.	Esti- mated IQs	Teachers' Ratings	Vine- land SQs
A	.71	.54	.61	A	.65	.52	.52
B	.87	.71	.72	B	.83	.73	.70
C	.87	.74	.75	C	.70	.69	.67
D	.83	.68	.70	D	.79	.67	.65
E	.84	.73	.71	E	.78	.67	.64

 $p < .001$ $p < .001$

Test III (N=102) (N=99) (N=102)			Test IV (N=102) (N=99) (N=102)				
Admin.	Esti- mated IQs	Teachers' Ratings	Vine- land SQs	Admin.	Esti- mated IQs	Teachers' Ratings	Vine- land SQs
A	.44	.47	.31*	A	.66	.66	.51
B	.84	.76	.70	B	.84	.75	.70
C	.83	.78	.64	C	.84	.78	.70
D	.82	.72	.68	D	.86	.76	.73
E	.81	.74	.68	E	.85	.75	.70

* $p < .01$ $p < .001$ all others $p < .001$

Test V	(N=102)	(N=99)	(N=102)
Admin.	Esti- mated IQs	Teachers' Ratings	Vine- land SQs
A	.77	.72	.70
B	.87	.76	.76
C	.84	.76	.72
D	.86	.76	.76
E	.87	.78	.75

 $p < .001$

the t s ranging from 2.66, ($p < .05$) to 7.99, ($p < .001$). The relationship between teachers' ratings and the learning scores was significantly greater than that between initial test performance and this criterion in 16 of the 20 comparisons, with t s from 2.23, ($p < .05$) to 5.04, ($p < .001$). The learning scores were also more significantly related to Vineland Social Quotients in 15 out of 20 cases, the t s ranging from 2.31, ($p < .05$) to 5.73, ($p < .001$).

Summary of Chapter

Test-retest comparisons showed that the scores obtained on various administrations of the tests were highly stable, and inter- and intratest comparisons demonstrated that they were also internally consistent. However, the learning scores, as compared with the initial scores, were always more stable and more consistent. For the group as a whole, highly significant differences between initial, posttraining, transfer, and retention performance were obtained on every test. The learning curves followed a highly consistent pattern, in accord with the different types of behavior which the various administrations were designed to measure.

It was found that the variables of sex and time in school did not significantly influence performance on any test. Age and diagnostic category exerted a limited but significant effect on some of the tests. In contrast, intelligence exerted a highly significant influence on every test, both alone and in interaction with the administrations. Intelligence was always more significantly related to learning performance than to initial performance, accounting for 18% to 55% of the variation in initial scores, and 62% to 76% of the variation in the learning scores. Of the learning scores, transfer and retention performance were the most sensitive measures

of differences in intelligence level.

The relationship between scores of the various administrations and teachers' ratings, estimated IQs, and Vineland Social Quotients were all significant. However, a significantly greater relationship was always found between learning scores and estimated IQs than between initial scores and this criterion. In the majority of instances, a significantly greater relationship was also found between learning performance and teachers' ratings than between initial scores and this criterion. In most instances, the learning scores were also more significantly related to Vineland Social Quotients.

IV

DISCUSSION

This chapter discusses the results of the pilot studies, the adequacy of the tests in meeting the criteria established for them, and the characteristics of the tests.

Results of the Pilot Studies

The immediate purpose of the pilot studies was to increase the adequacy of the tests in meeting the criteria set for them. During this preliminary experimentation, a number of ancillary findings were obtained, which suggest interesting possibilities for further research in two main directions. The first is the application of the special techniques used in improving the materials and training procedures in the test situation to curricular materials and training methods in the classroom. The second is the extension of the method of testing used in this research to other areas.

Findings Related to Materials and Methods for Training

Chief among the problems which arose in the development of the battery were the following: increasing motivation for learning, improving the training procedures used in the training periods, and minimizing perseveration. A number of successive changes were introduced into the pilot studies on an empirical basis, which progressively increased the efficiency of the tests in these respects. The specific set of variables responsible for these improvements cannot be identified at this time. However, the findings summarized below, if investigated further, may have direct applicability to training in the classroom. The findings are based largely on data provided in Appendices I and II.

1. Simple sensory rewards, such as the sound of bells or chimes, and the turning on of lights, were the most satisfactory motivating devices.

2. These rewards could be built into the materials or supplied by the examiner as the child performed. A combination of both was the most successful.

3. The positive effect of these devices lasted for a relatively short period. Thereafter, introduction of another such reward was necessary to reawaken interest and maintain it for another brief period.

4. The factor of novelty in the stimulus appeared to be more important than changing the sense modality involved; i.e., substituting one sound for another was as effective as changing from an auditory to a visual stimulus.

5. The use of external rewards, such as offers of candy or toys, were often helpful in establishing rapport, but did not seem to be a significant factor in motivating the child's performance as such.

6. The usual testing practices of smiling, nodding approval, supplying quiet, steady reassurance, and so on, were highly unsatisfactory.

7. An extremely active, overt, and constant participation on the part of the examiner was a highly facilitating factor, especially in the form of clapping, laughing, singing, duplicating the sounds of the test materials verbally, patting the child, and so on.

8. Training to the correct responses only was considerably more effective than training to both selection of correct responses and avoidance of incorrect ones.

9. When it was necessary in the pretest demonstrations to show

the subject that certain materials were nonfunctioning, it was essential to reinforce the correct response immediately thereafter.

10. The most suitable training methods were those in which all procedures were repeated without variation. Whether this would be so in longer training periods is not known. However, it can be said that the most effective combination for facilitating learning in the short training sessions was that of constant changes in the materials, without the introduction of change into the training methods.

11. Perseveration was minimized when the pattern of the required responses was not changed, and only the number of the required responses increased. The introduction of even minimal changes into the pattern tended to result in an increase of perseveration.

12. Items which required choosing among materials arranged so that increments in size or degree of brightness between adjacent material was constant tended to increase perseveration.

13. The tendency to persevere was minimized by placing the most dissimilar objects next to one another, and also arranging the items so that the child was never required to choose adjacent objects successively.

14. Perseveration was apt to occur if training began with a more complex item in a test which required a pattern of responses. Under those circumstances, the subject was often unable to return to the simpler responses which he had previously given correctly. It was therefore more successful always to begin training with the simplest item, regardless of the items the subject had previously passed.

Findings Related to Extending the Testing Method to Other Areas

1. Several tests in the experimental battery, and some of the earlier forms of the final tests, were rejected as too difficult for

this population. Some of these were tried out experimentally with ten moderately retarded children, aged 5-8 to 11-1, with IQs on the Revised Stanford Binet Scale ranging from 51 to 74. The number of subjects was too small to warrant generalizations. However, the obtained results suggest that these tests may be suitable for evaluating learning abilities in moderately retarded children, providing a basis for an upward extension of the tests.

2. A number of tests were rejected as too easy for discrimination purposes with the population tested here. These were tried out experimentally with eight severely retarded children, aged 3-6 to 4-10. Though again the number of subjects was small, the results suggested that these tests may be useful in measuring learning abilities in severely retarded children below the age range used in the present study, and may serve as a basis for a future age scale.

3. Of considerable interest are the results obtained by ten children medically diagnosed as schizophrenic. The final battery was used for these subjects. They were among a group of 24 children at a new school, who had already been given psychological evaluations, but not medical examinations. Since this group was available in early September, when subjects were scarce, all of the children were tested, and results set aside pending the medical reports. The results showed that eleven subjects had performed markedly better on all tests than was usual with subjects at comparable mental levels. When the medical reports arrived, ten of these eleven subjects were found to be diagnosed as schizophrenic, and they were therefore excluded from the sample. However, it is possible that the use of learning tests may provide a basis for differentiating the so-called "pseudo-retardate" from the mentally retarded child functioning at a comparable mental level.

4. Results obtained on tests which were rejected as unsuitable for brain-injured subjects are also of special interest. Preliminary analysis of the limited results obtained suggest that the brain-injured subjects were more sharply differentiated from the non-brain-injured subjects by the learning scores than by their initial test performance. It may be that learning responses, as obtained by tests such as these, are better indicators of the presence of brain-injury than are single administrations of a test or tests.

Summary

Findings obtained from the pilot studies suggest that the various techniques used for increasing motivation for learning, improving the training procedures, and minimizing perseveration which were used in the development of the tests may have direct applicability to materials and methods for classroom training. The results of the pilot studies also suggest the suitability of the method used in this study for research with other age levels and other retarded groups, and for purposes of differential diagnosis.

Adequacy of the Tests in Meeting the Criteria

The first consideration in discussing the results obtained by the final battery is their bearing on the adequacy of the tests in meeting the criteria established for them. Criterion 1 required that tests be within the abilities of the subjects. Apparently the tests were within their abilities, since, as a group, the subjects obtained significantly higher scores after training than before it, and were also able to transfer and retain the learning to a significant extent. It can also be said that, within the range of abilities studied, the tests were able to discriminate among the subjects, since those at

different levels of ability produced significantly different scores.

Criterion 2 required that the tests depend chiefly on abilities sufficiently well developed at the age levels of the subjects. The obtained results demonstrated a limited but significant effect of age on some of the tests. Further research may indicate that different norms are necessary for different age levels. However, the tests themselves were suitable for the age range tested, since learning performance was obtained for all of the age levels.

Criterion 3 demanded that the tests be suitable for brain-injured subjects. While the results showed some inferiority in performance of the brain-injured subjects as compared to that of non-brain-injured groups on two of the tests, significant learning responses were obtained by the brain-injured subjects on all of them. Though the effect of this variable on test performance requires further study, the obtained results showed that the battery was essentially suitable for these children.

That the tests were interesting to the subjects and short enough to hold their interest, (Criteria 4 and 5), was shown by the cooperation of the children. Their behavior in the test situation clearly pointed to the adequacy of the tests in these respects. The quantitative results supported these qualitative observations. Both the intertest and intratest comparisons were highly consistent, as were the test-retest comparisons. A highly consistent pattern of learning curves was found on all of the tests. Neither this degree of consistency, nor the highly significant learning scores which were obtained on all tests, would be likely to occur in the face of lack or loss of interest.

Adequacy in meeting Criterion 6, which required minimal depen-

dence on past experience, can only be inferred. Probably the most directly relevant finding was that the variable of time in school did not significantly influence performance on any administration of any test. The range of time in school extended from 0 to 36 months, or up to 5 years, since only 9 months were counted for each year in estimating this variable, vacation months being excluded. Differential experiences are necessarily involved in a five-year range of school attendance. Further, these are precisely the kinds of experiences which would be most likely to influence performance on tests such as these, especially since materials and procedures are not unlike those used in the training schools. It can therefore be said that precisely those kinds of experiences which might reasonably be expected to exert a differential influence on test performance failed to do so.

Criteria 7, 8, and 9 were met in the construction of the materials, and the procedures by which the tests were administered. The tests did not depend on verbal instructions, and no verbal or manipulative responses on the part of the subject were required. Children with no speech, and those with severe physical handicaps, were able to learn in all of the test situations. The use of an achromatic scale served to avoid difficulties in color perception.

Summary

The results indicate that the tests met the criteria established for them. The tests were appropriate to the abilities of the subjects, aroused and held the interest of the children, and were suitable for subjects with severe verbal and physical handicaps.

Characteristics of the Tests

This section discusses the reliability and internal consistency of the tests, comparisons of scores obtained on various administrations, the effects of population variables on test performance, and the validity of the tests.

Reliability and Internal Consistency of the Tests

Test-retest comparisons demonstrated a high degree of stability over time for all administrations of the tests. However, learning performance was found to be more stable than initial test performance with coefficients of .90 or better in every case. The learning scores were also characterized by a greater degree of internal consistency than were the pretraining scores, both across and within the tests.

In brief, learning performance was more closely related than were initial scores to later trials, to other learning trials on the same test, and to the same trials on other tests. This suggests that pretraining performance was responsive to more specific factors, probably including previous experience. The greater stability and consistency of the learning scores suggests that a more general factor was measured.

Comparisons of Scores Obtained on Various Administrations

The learning curves obtained on each test followed a highly consistent pattern, and one which would be logically anticipated in view of the different types of behavior the administrations were designed to measure. On every test, the learning scores obtained by the group as a whole were significantly greater than initial performance, clearly demonstrating that the group had learned to a

significant extent. The largest learning gains were always obtained on posttraining performance, where the effect of the training was most direct. Transfer scores were always significantly lower than posttraining scores and significantly higher than initial performance, indicating that the group was able to transfer some of the training to other situations. Retention scores were always lower than posttraining scores and higher than pretraining scores, showing that the group had retained a significant amount of the learning.

It is interesting to note that no reinforcement of the learning was required to obtain a significant amount of retention within an hour of the original training. With reinforcement limited to repetitions of the original instructions and pretest demonstrations, a significant amount of retention was obtained a week later. Again with the same reinforcement, significant retention was also obtained by four of the five tests after an interval of seven weeks.

The Effects of Population Variables on Test Performance

No significant effect of the factors of sex or time spent in school was found on any administration of any of the tests. Apparently, the battery was suitable for boys or girls, with school experience ranging from none through five years.

Age was found to influence performance on some of the tests, which might be expected in an age range of almost seven years. In the obtained results, age was not found to influence pretraining and learning performance differentially; only the youngest subjects were penalized and only on tests as a whole; and no significant effect of age was obtained on two of the tests. However, this restricted effect may be due to the size of the sample. A larger population is needed, since there are some indications that age norms may be necessary.

A larger sample is also needed for more definitive study of the effect of diagnostic category on test performance. For this sample, a limited inferiority was found for the brain-injured subjects on two of the tests, and no significant influence of the factor of diagnostic category was obtained on the rest of the battery. However, the results obtained were a function of the levels used for determining the effect of this variable, and these groups were not suitable for definitive study. Although the brain-injured subjects were grouped together, brain-injury is a heterogeneous phenomenon. Additional research is also needed in connection with the diagnostic categories included in the miscellaneous group.

Intelligence, (as measured by estimated IQs), was the only factor which was found to be highly related to performance on every test. The learning scores were always more sensitive measures of differences in intelligence level than was pretraining performance. The factor of intelligence accounted for 18% to 55% of the variation obtained on initial test performance, and for 62% to 76% of the variation in the learning scores.

Intelligence not only exerted a differential influence on pretraining and learning performance, but also on the different types of learning behavior which were investigated; i.e., posttraining gains, transfer, and retention. The low subjects made significant post-training gains in most cases, but did not transfer the training or retain the learning to a significant extent. The middle group learned, transferred the training, and retained the learning to a significant extent on all tests, while the high group always made the largest gains in all of the learning areas. The fact that the deficiencies of the low group were most clearly shown on transfer

and retention performance indicates that these scores were the most sensitive of the learning measures to differences in level of intelligence.

Validity

The obtained results consistently attest to the construct validity of the tests. The evidence on its behalf began with the demonstrated adequacy of the tests in meeting the criteria which were established for them. The battery was found to be suitable to the abilities of the subjects, to exert minimal penalty for their verbal, manipulative, and sensory lacks, and to be capable of differentiating levels of ability within the group. These were preconditions of validity.

The basis for construct validity lay in the learning abilities which were studied. The ability to learn from training, transfer the training, and retain the learning are virtually a definition of educability itself. All education depends on them, and all education is directed toward their development.

If these were indeed the abilities which the tests measured, the following pattern would be the logical outcome: The lowest scores would obviously be found on initial performance, and the learning scores would be significantly higher. Since training was given between the first and second administrations, the greatest differences would be obtained between these scores. Transfer performance would be both significantly below posttraining performance, and significantly higher than initial scores. Retention scores would be significantly greater than pretraining performance. This was precisely the curve which was obtained on every test.

The validity of the constructs would also demand the stability

of the learning performance over time, and, if more general learning abilities were tapped, a high degree of consistency among the learning scores. The results demonstrate that these conditions were met. Further, if learning performance were an adequate measure of actual differences in abilities, the magnitude of the learning scores would be different at different levels of ability, and would increase as the level of ability increased. Finally, it would be necessary that the learning scores be significantly more sensitive to these differences than pretraining scores. These results were obtained for every test.

Establishing the predictive power of a test requires a representative sample, intensive study, and considerable time. The power of the tests to predict specifically the ability of the child to learn in the classroom situation remains to be demonstrated, and would be required if the battery were to be used for purposes of selection or for making actual educational decisions. Meanwhile, the obtained results give much indirect evidence that the battery would be useful in such cases. These are found in the comparisons of the test scores with the external criteria.

Before considering the relationships between the test scores and the external criteria, a word concerning the criteria themselves is in order. Teachers' ratings constitute the closest approach to a criterion of educability, in that they reflect the actual experience of teachers with the child's classroom learning over a period of time. The measure was a reliable one, as was shown by the highly significant correlation obtained by comparing the ratings of two teachers on the same children, each without knowledge of the other's rating.

A highly significant relationship was also found between teachers' ratings and psychologists' estimated IQs, although teachers and psychologists were unaware of each others' estimates. Though this may show that the measures supported each other, it must be remembered that the psychologists' records were available to the teachers, and the extent to which their ratings were influenced by this knowledge cannot be determined. However, it can be said that, while making the actual ratings, the teachers did review their own longitudinal records of the child's behavior, but did not refer to his psychological evaluations at that time. It can also be said that, though the ratings of the teachers were significantly related to all of the other criteria, they were not too closely related to preclude independent judgment.

Psychologists' estimated IQs, on the other hand, were very closely related to the scores of standard tests, correlating .92 with Binet IQs where obtainable, and .81 with Vineland Social Quotients. It has been shown earlier that estimated IQs were the most appropriate measure of intellectual functioning which was available for all of the subjects. Since some amount of subjectivity entered into this measure, the highly significant relationship between this estimate and the scores of standard tests was desirable. The reliability of estimated IQs was demonstrated by the highly significant correlation between two such estimates obtained for the same children, each made independently of the other.

The Vineland Social Maturity Scale is a measure of social competence. Although it is not a measure of learning ability per se, the concept of social competence is relevant to the training of these children in that the curriculum emphasizes the development of traits which it reflects.

The comparisons between the test scores and the external criteria gave a partial demonstration of predictive validity. All of the scores were significantly related to all of the criteria. However, a marked difference in the degree of the relationship was found for pretraining performance on the one hand, and learning performance on the other.

In all cases, the learning scores yielded significantly higher coefficients when compared with psychologists' estimated IQs than did pretraining performance. In the majority of instances, the relationship between the learning scores and teachers' ratings was significantly greater than that between pretraining scores and this criterion. In most cases, the learning scores were also more closely related than initial scores to Vineland Social Quotients. Thus, the learning scores were more closely related than pretraining scores to classroom learning, intelligence, and social competence, as measured by teachers' ratings, estimated IQs, and Vineland Social Quotients, respectively.

Summary of Chapter

Results of the pilot studies suggest a number of possibilities for future research, both in improving curricular materials and training methods in the classroom, and in extending the method of testing used in this study to other areas. The results based on the final battery show that the tests were appropriate to the abilities of the subjects with severe verbal and physical handicaps. The various administrations of the tests were shown to be both stable and internally consistent. However, learning performance demonstrated a greater degree of stability over time, and of internal consistency both within and across tests, than did initial test performance.

The group as a whole made highly significant learning gains on every test. Within the learning scores, significant differences were also found between posttraining, transfer, and retention performance. The learning curves obtained for every test were highly consistent, and in close accord with the pattern which would logically be anticipated from the types of behavior measured. They were also stable over time.

The variables of sex and time in school did not significantly affect test performance. The factors of age and diagnostic category had a limited influence on some of the tests. Intelligence was the only variable found to affect performance to a highly significant extent on every test, the relationship between intelligence and learning performance being significantly greater than that between intelligence and initial test scores. Among the learning scores, transfer and retention were found to be the most sensitive measures of differences in intelligence level.

Evidences of construct validity were found in the shape of the learning curves obtained, and their consistency and stability; the greater stability and consistency of the learning scores as opposed to initial scores; and the significantly greater relationship between intelligence level and learning performance than between pretraining performance and intelligence level.

Partial demonstrations of predictive validity were also obtained. All scores were significantly related to teachers' estimates, psychologists' estimated IQs, and Vineland Social Quotients. However, the learning scores were more closely related than pretraining scores to all of the criteria.

SUMMARY AND CONCLUSIONS

Summary

The purpose of this study was to investigate a method for obtaining an objective, quantified, predictive measure of educability for the severely mentally retarded child. The method was designed to measure abilities on which education depends; namely, the ability to learn under training, to transfer the training, and to retain the learning. The subject was first given a battery of tests without training. He was then trained to the correct responses according to standardized training procedures, and retested so that learning gains could be measured. Different forms of the tests were used to evaluate transfer ability. Retention scores were obtained by retesting the subject at various later dates.

Criteria were established before the battery was constructed, which required that the tests be appropriate to the abilities of the children; capable of arousing and maintaining their interest; and suited to their verbal, motor, sensory, and experiential handicaps. Areas selected for study were imitative abilities, memory, and gross shape, brightness, and size discriminations. A preliminary battery consisting of eleven tests was constructed, and administered to 115 retarded children, in a series of pilot studies whose purpose was to bring the tests into increasing accord with the criteria.

In establishing the upper limit of the range of intelligence to be used for the sample for the final battery, the following criteria were used. An IQ below 50, obtained on the Revised Stanford-Binet Intelligence Scale administered within the year, was used for all

subjects to whom this test was applicable. Where the test was inapplicable, an IQ under 50 on other standard tests of intelligence, given within the year, was accepted. An estimated IQ below 50 was used for subjects who failed to respond in a measurable way to any standard intelligence test. Estimated IQs were obtained from the psychologist on the staff or clinic the child attended, who based them primarily on quantitative and qualitative analyses of the child's psychological evaluations, which always included a Vineland Social Quotient, at least fragmentary results of standard intelligence tests, and scores from these tests if obtainable.

The final battery was administered individually to 114 severely mentally retarded subjects, living in the community, and attending special classes and/or clinics for mentally retarded children. Their estimated IQs ranged from 8 to 55, mean 36.18, standard deviation 12.01. Their ages ranged from 61 to 143 months, mean 95.16 months, standard deviation 23.48. They had attended school from 0 to 36 months, mean 11.42 months, standard deviation 10.48. They were medically stated to suffer from various pathological physical conditions known to produce mental retardation, such as Mongolism, cerebral palsy, and microcephaly. There were 69 boys and 45 girls.

The final battery, which consisted of five tests, included a test of imitative abilities and memory, a size discrimination test, a shape discrimination test, and two tests of brightness discrimination. The battery was administered nine times. The first three administrations yielded pretraining, posttraining, and transfer scores respectively. The fourth and fifth administrations produced retention scores. The last four administrations were repetitions of the first

four, given at a later date. All of the subjects received the first four administrations of the battery. Of the original subjects, 102 were still available for the fifth administration, and 36 for the last four.

The results demonstrated that the scores obtained by the various administrations of the tests were highly reliable, as shown by test-retest comparisons. A high degree of internal consistence, as shown by inter- and intratest comparisons, was also found for all tests. However, as compared with initial test performance, learning performance, (i.e., posttraining, transfer, and retention scores), were found to be more stable, with coefficients of .90 or better in every case. The learning scores were also found to be more closely related than initial scores to other learning trials on the same test, and to corresponding trials on different tests, suggesting that initial performance was more responsive to specific factors, and that a more general factor was measured by the learning scores.

Comparisons of scores obtained on the various administrations demonstrated that on every test the group as a whole had learned, transferred the training, and retained the learning to a highly significant extent. The shape of the learning curves was highly consistent and in close accord with the various types of behavior the administrations were designed to measure. The greatest differences were always obtained between pre- and posttraining scores, where the effect of the training was most direct. Transfer scores were always significantly lower than posttraining scores and significantly higher than pretraining performance, indicating that some of the training was transferred to other situations. Retention scores were always lower

than posttraining scores and significantly higher than initial scores, showing that the learning was retained to a significant extent.

The variables of sex, time spent in school, age, diagnostic category, and intelligence (as measured by estimated IQs) were together found to account for 23% to 61% of the variation in initial scores, and for 69% to 80% of the variation in learning performance. Considered individually, sex and time in school were found to have no significant influence on test performance. Age and diagnostic category had a limited effect on performance on some of the tests. Intelligence exerted a highly significant influence on the scores of every test, both as a main effect and in interaction with the administrations. Intelligence was always more significantly related to learning scores than to pretraining scores, accounting for 18% to 55% of the variation in initial performance, and 62% to 76% of the variation in learning performance. Among the learning scores, transfer and retention scores were found to be the most sensitive measures of differences in intelligence level.

The obtained results consistently attested to the construct validity of the tests. Evidence was found in the results which showed that the learning scores, as opposed to the initial scores, were more stable across time and more consistent both within and across the tests; in the highly consistent learning curves in close accord with the constructs; and in the significantly greater relationship between learning performance and intelligence level than between intelligence level and initial performance.

Comparisons of the scores obtained by the various administrations of each test with teachers' ratings, psychologists' estimated IQs,

and Vineland Social Quotients gave further evidence of construct validity, and also partial demonstration of predictive validity. All of the scores were found to be significantly related to the three criteria. However, learning performance, as measured by the tests, was found to be more significantly related than initial scores to classroom learning, intelligence, and social competence, as measured by teachers' ratings, estimated IQs, and Vineland Social Quotients, respectively.

A number of ancillary findings were also obtained which suggest that the materials and methods developed for the tests have direct applicability to curricular materials and training methods in the classroom, in terms of increasing motivation, minimizing perseveration, and improving training procedures. The preliminary findings also suggest that the method of testing used in this study is applicable to other groups, including moderately retarded and younger severely retarded children.

Conclusions

1. The factors of sex and time spent in school did not affect test performance to a significant extent.
2. Age and diagnostic category had a limited influence on the scores of some of the tests.
3. Intelligence exerted a highly significant effect on the scores of every test.
4. The relationship between learning performance and intelligence level was significantly greater on every test than that between intelligence level and initial performance.
5. Among the learning scores, transfer and retention performance were the most sensitive measures of intelligence level.

6. Learning performance was more closely related to teachers' ratings, psychologists' estimated IQs, and to Vineland Social Quotients than was initial performance.
7. The tests were reliable.
8. Good evidence on behalf of construct validity, and partial demonstration of predictive validity, were obtained.

The ancillary findings suggest:

1. The applicability of the materials and training methods developed for the tests to curricular materials and training methods in the classroom.
2. The suitability of the method of testing used in this study to other groups of children, including the moderately retarded, and the younger severely retarded.

Helen Schucman,
Principal Investigator.

Bibliography

- Barnett, C. D. and Cantor, G. N., "Discrimination Set in Defectives," Amer. J. Ment. Defic., Vol. 62 (1957), p. 334.
- Cox, D. R., Planning of Experiments, New York: John Wiley and Sons, 1958.
- Crafts, L. W., Schneirla, T. C., Robinson, E. E. and Gilbert, R. W., Recent Experiments in Psychology, New York: Appleton-Century-Crofts, 1948.
- Edwards, A. L. Experimental Design in Psychological Research, New York: Rinehart and Co., 1956.
- Guilford, J. P., Psychometric Methods, New York: McGraw-Hill, 1954.
- Harlow, H. F., "The Formation of Learning Sets," Psych. Review, Vol. 54 (1949), p. 51.
- Hebb, D. O., The Organization of Behavior: A Neuropsychological Theory, New York: John Wiley and Sons, 1949.
- Hilgard, E. R., Theories of Learning, New York: Appleton-Century-Crofts, 1948.
- House, B., Zeaman, D., Orlando, R. and Fischer, W., Learning and Transfer in Mental Defectives, Univ. of Conn. 1957.
- McLachlan, D. G., "Emotional Aspects of Backward Children," Amer. J. Ment. Defic., Vol. 60 (1955), p. 323.
- McNemar, Q., Psychological Statistics, New York: John Wiley and Sons, 1955.
- Penrose, L. S., Mental Defect, New York: Farrar and Rinehart, 1934.
- Porteus, S. D., Studies in Mental Deviation, New Jersey: Vineland Training School, 1923.
- Peters, C. C. and Van Voorhis, W. R., Statistical Procedures and Their Mathematical Bases, New York: McGraw-Hill, 1940.
- Pressey, S. L. and Robinson, F. P., Psychology and the New Education, New York: Harper and Bros., 1944.
- Schucman, H., "A Method for Measuring Educability in Severely Mentally Retarded Children: A Preliminary Study," Training School Bull., Vol. 54, Nos. 3 and 4, (1957 and 1958) and Vol. 55, No. 1, (1958).
- "Technical Recommendations for Psychological Tests and Diagnostic Techniques," Supplement to Psych. Bull., Vol. 51, No. 2, Part 2, (1954).

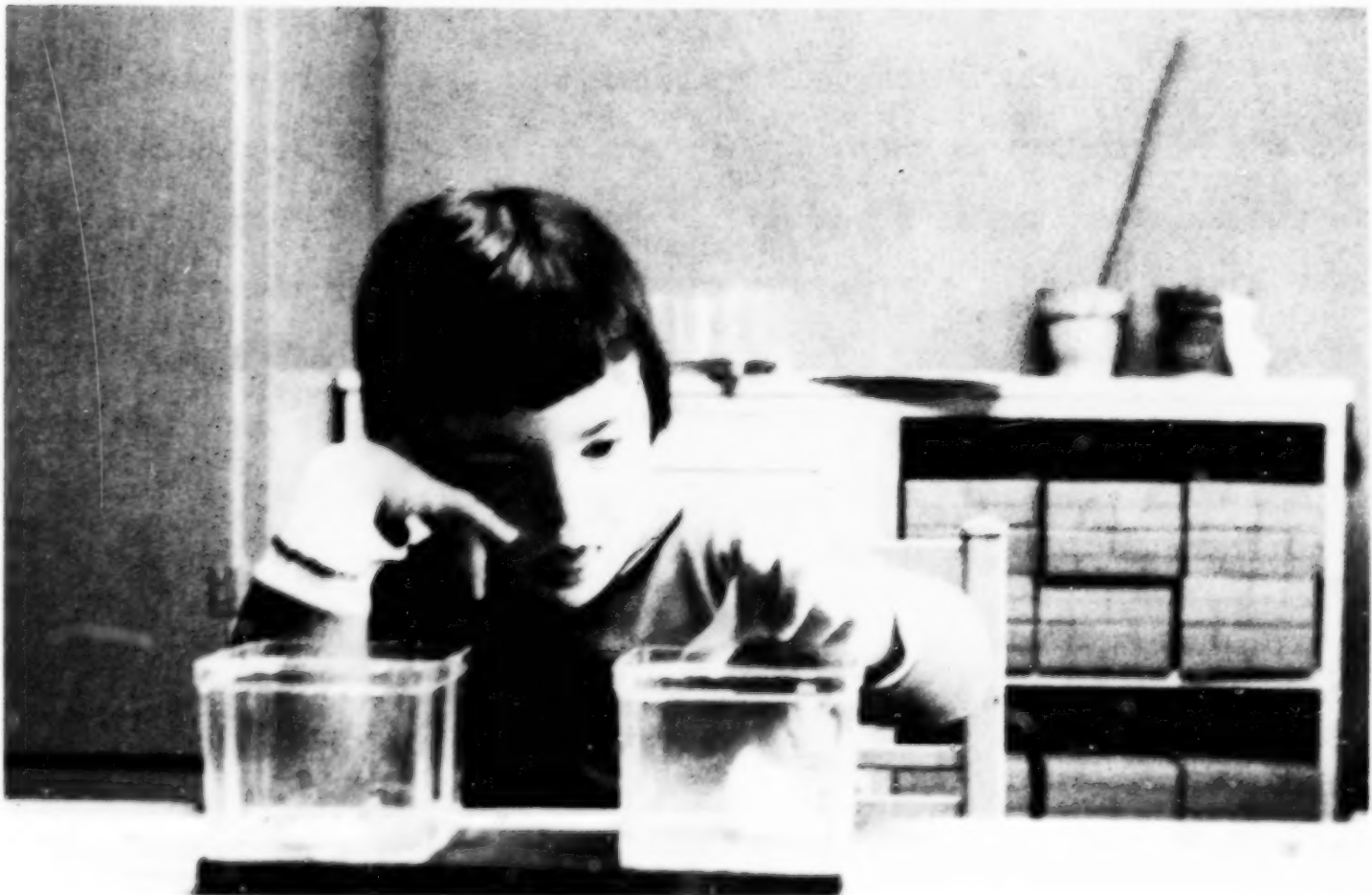
Walker, H. M. and Lev, J. Statistical Inference, New York: Henry Holt and Co., 1953.

Webb, L. W., "Transfer of Training," in C. E. Skinner (ed.), Educational Psychology, New York: Prentice-Hall, 1950.

Appendix I

Appendix I
Materials and Procedures for
Administering the Tests

I. The Two-Buttons Test



Materials

Two clear plastic boxes, 3 and 1/2 inches high, with sides 3 and 1/2 inches along the bottom and 4 inches along the top. The boxes were attached, 3 and 1/2 inches apart, to a black board 10 and 1/2 inches long and 3 and 1/2 inches wide.

Eight round white buttons, 1 and 3/4 inches in diameter.

Testing Procedures

There were no special pretest procedures for this test, since demonstrations preceded each item. For each item, the subject was required to indicate the box into which each button presented should be placed, in imitation of the examiner's demonstration. If necessary, the examiner picked up the buttons for the subject, and dropped them into the boxes at his indication.

The test consisted of six items, which are shown on the appended scoring sheet. Each "L" on the scoring blank represents one button to be dropped into the box at the subject's left. Each "R" stands for one button to be put into the box at his right. The first three items required the correct placement of single buttons, one to be dropped into the box at the left, the next into the box at the right, and the third into the box at the left. The fourth item consisted of two buttons, the fifth of three, and the sixth of four. The buttons were dropped into the boxes in alternating order, always beginning at the left.

Boxes were placed about two inches from the edge of the table. Buttons were presented approximately an inch from the edge of the table, and three inches apart. Before each item was presented, the examiner demonstrated the correct response, dropping the buttons into the boxes from a height of approximately a foot to ensure a sharp click. As the buttons struck the box, the examiner said, "Ping!" in a singing tone, drawing out the final "ng." After dropping each button, the examiner tapped the box into which it had been dropped, and said, "in here." Buttons used in the demonstrations were not removed until the subject's response was secured.

After each demonstration, the examiner placed the number of buttons required for the item before the subject, and said, "You do it," pointing from the subject to the buttons, and then to the boxes, midway between them. All buttons were removed before the next demonstration. The examiner said, "Ping!" whenever the subject gave a correct response. If the response was wrong, the examiner corrected it, saying, "Ping!" as the buttons were placed correctly.

The first three items were always given. If all three were failed, the test was discontinued and training was started. Testing was also discontinued if the subject failed the second, third, and fourth items. If he failed the fourth and fifth items, the sixth was omitted.

Training Procedures

Training began with the subject's first failure unless it had occurred on one of the first three items, in which case training began with the first item. Demonstrations were made as before. The examiner then presented the buttons to the subject, saying "in here" for each button, pointing to the correct box, and tapping it until the response was made. Again, "Ping!" accompanied all correct or corrected responses.

The subject was trained three times on each item, unless he achieved two consecutive successes. In that event, the third training was omitted. Training was discontinued if there was complete failure on the first three items. Training was also stopped if the second, third, and fourth items were failed. Training on the sixth item was omitted if no successes were achieved on the fourth and fifth.

The Transfer Form

The original items were those used in the transfer situation. However, the demonstrations were omitted. In advance of each presentation, the examiner pointed to the correct box or boxes, saying, "Ping! in here," or "Ping! in here" and "Ping! in here," and so on. As before, correct or corrected responses were accompanied by, "Ping!"

TWO BUTTONS TEST

Administration A

(1) L (2) R (3) L (4) L-R (5) L-R-L (6) L-R-L-R Total _____
.....

Training

L	R	L	L-R	L-R-L	L-R-L-R
L	R	L	L-R	L-R-L	L-R-L-R
L	R	L	L-R	L-R-L	L-R-L-R

.....

Administration B

(1) L (2) R (3) L (4) L-R (5) L-R-L (6) L-R-L-R Total _____
.....

Administration C

(1) L (2) R (3) L (4) L-R (5) L-R-L (6) L-R-L-R Total _____
.....
.....

Administration D

(1) L (2) R (3) L (4) L-R (5) L-R-L (6) L-R-L-R Total _____
.....

Administration E

(1) L (2) R (3) L (4) L-R (5) L-R-L (6) L-R-L-R Total _____
.....

Comments:

II. The Buzz Test



Materials

A gray board, 23 inches long and 7 and 1/2 inches wide, attached at an angle of approximately 45 degrees to a base and a back board which joined at a 90 degree angle. Four black discs, 1 inch in thickness, and with diameters of 3, 6, 4, and 5 inches respectively, were attached to the board in that order, with equal distances between them. Thus, beginning at the left, the smallest disc was first, the largest second, the next to the smallest third, and the next to the largest fourth. The sample discs on the board were wired so that, when a disc of corresponding size was placed over one of the samples, a continuous buzzing

Al-7

sound was produced. Two blunt nails projected from the board below each sample, to hold a disc of comparable size in place.

One set of separate black discs and one set of separate white discs, one disc in each set corresponding in size to each of the sample discs on the board.

Pretest Procedures

The board was placed about four inches from the edge of the table. The examiner held the smallest black disc at the center of the board, about one inch from the edge of the table, at an angle corresponding to the tilt of the board. After holding the disc in this position for several seconds, the examiner placed it over the corresponding sample on the board, saying, "This goes here." After allowing the resulting buzz to continue for a few seconds, the examiner returned the disc to the center of the board, holding it as before. Pointing from the subject to the disc, and then to the middle of the board, the examiner said "You do it." The examiner placed the disc at the subject's indication, if necessary. One demonstration and one unscored response were allowed for each of the black discs. All errors were corrected.

Testing Procedures

For each item, the examiner held a black disc in the position described above, asking each time, "Where does this go?" and pointing from the disc to the board. The subject was required to indicate a choice of one of the samples, the examiner placing the disc for him, if necessary. Errors were corrected before continuing to the next item.

There were six items, each involving matching one black disc to the corresponding sample. The scoring sheet shows the order in which the discs were presented. The number in parenthesis on the scoring sheet refers to the number of the item. The number which follows shows the position on the board of the sample corresponding in size to the disc presented, i.e., "2" represents the largest sample, which was second on the board. All of the items were given, unless the subject failed the first four, in which case the last two were omitted.

Training Procedures

Training always began with the first item. The examiner began by demonstrating the response, saying, "This goes here," producing the buzz, and then removing the disc and presenting it to the subject, as before. The examiner pointed to the correct sample, and continued to tap it, repeating, "Here" until the response was made. If errors still occurred, they were corrected by the examiner.

A special order of the items was used in training on this test, in order to minimize the tendency to perseveration. Instead of the usual procedure of completing training on each item before beginning training on the next, the subject was trained once on all items, in the order in which they were presented in the test, beginning with item 1. The entire series was repeated three times, unless no errors were made in the course of the first two series.

The Transfer Form

Using the same items, and following the original procedures, the white discs were presented to the subject, to be matched to the black samples of corresponding sizes on the board.

BUZZ TEST

Administration A

(1) 1 (2) 3 (3) 2 (4) 4 (5) 1 (6) 4 Total _____
.....

Training

1 3 2 4 1 4 --- 1 3 2 4 1 4 --- 1 3 2 4 1 4
.....

Administration B

(1) 1 (2) 3 (3) 2 (4) 4 (5) 1 (6) 4 Total _____
.....

Administration C

(1) 1 (2) 3 (3) 2 (4) 4 (5) 1 (6) 4 Total _____
.....
.....

Administration D

(1) 1 (2) 3 (3) 2 (4) 4 (5) 1 (6) 4 Total _____
.....

Administration E

(1) 1 (2) 3 (3) 2 (4) 4 (5) 1 (6) 4 Total _____
.....

Comments:

III. The Music Box Test



Materials

Five round music boxes, 2 inches high and 3 inches in diameter. A handle, which could be turned in either direction, projected 2 inches from the center of the box. The shaft of the handle extended vertically from the box for a distance of one inch, and was then bent at a right angle, continuing horizontally for another inch, and ending in an upright wooden cylinder, a half inch in diameter and one inch long. One box was painted black, one white, one gray, one black and gray, and one gray and white. Only the gray box functioned, playing a repeated series of four tones.

Al-11

Five rectangular music boxes, 2 and 1/2 inches high, 4 and 1/2 inches long, and 2 and 3/4 inches wide, in other respects like the round boxes.

Pretest Procedures

The round gray box was placed before the subject, about an inch from the edge of the table, with the handle toward him. The examiner turned the handle for some three seconds, saying, "music - music - music." The examiner then said to the subject, "You make music," pointing from the subject to the gray music box. If necessary, the examiner placed the child's hand over or against the handle and turned it for him, repeating, "music - music - music."

The gray box was left on the table, but was pushed aside, some six inches to the subject's left. The round black and gray box was presented in center position, its black side toward the gray box. The examiner turned the handle about three seconds, shaking his head and saying, "no music, - no, - no music." The subject was then asked to turn the handle, as before, the examiner repeating, "no music" as he responded.

The black and gray box was pushed to the subject's right, and the round gray and white box was presented, with the white side toward the gray box. The demonstration and response were repeated. The box was then pushed to the subject's left, and the gray box was presented again. The examiner played it, again for about three seconds, saying, "here's the music," and asking the child to play it, as before. The boxes were removed before the test began.

Testing Procedures

Boxes were presented about four inches apart, the handles turned toward the subject. At each presentation, the examiner said, "You make Music," pointing from the child to the music boxes. The subject was required to indicate his choice, the examiner turning the handle of the box he selected, if necessary. If the subject's response was wrong, the examiner pointed to the gray box, turned the handle, and said, "here's the music."

The shifts in the positions of the boxes were made with the subject watching. However, to prevent uninterrupted visual fixation on the gray box, the subject's attention was briefly diverted during each shift. To keep the interruption as standard as possible, the examiner said, "Look at me, (subject's name)," and then redirected the child's attention to the table. This required two or three seconds.

The scoring sheet shows the number of boxes used for each of the six items, and the order in which they were presented. The letter "G" on the sheet represents the gray box, "B" the black, "W" the white, "B/G" the black and gray, and "W/G" the white and gray. Two boxes were used for each of the first two items. Three were used for the third, and four for the fourth. Five boxes were used for the fifth and sixth items. The two-toned boxes were not introduced into the test until the fourth item. One of them was used in the fourth item, and both in the fifth and sixth. In the fifth item only one of the two-toned boxes was placed next to the gray box. In the sixth item, the gray box was presented with a two-toned box on each side. The two-toned boxes were always presented with the gray side away from the gray box. The first three

items were always given. The test was discontinued if these three items, or the second, third, and fourth items were failed. The sixth item was not given if the subject failed the fourth and fifth.

Training Procedures

If the subject's first failure occurred on one of the first two items, training began with the first. Otherwise, it began at his first failure. The examiner presented the item, turned the handle of the gray box, and said, "music - music - music." Then, pointing from the subject to the gray box, the examiner said, "You make music," tapping the gray box, and repeating, "Here's the music" until the response was secured. If necessary, the examiner placed the handle of the gray box into or against the subject's hand, turning the handle for him, and saying, "music - here's the music." No interruptions were made in the shifts of materials during training.

The subject was trained three times on each item, or through two consecutive successes. Training was always given on the first three items. It was discontinued if these were failed completely, or if complete failure occurred in training on the second, third, and fourth items. Training was omitted for the sixth item, if no successes were achieved on the fourth and fifth.

The Transfer Form

The rectangular music boxes were used in place of the round ones. The original testing procedures were followed, and the original test items were used. The round set was pushed to the back of the testing table, but was not removed, in an attempt to emphasize the fact that different boxes were being presented.

MUSIC BOX TEST

Administration A

(1) G-B (2) W-G (3) G-W-B (4) B-W-G-W/G (5) W/G-G-W-B-B/G
(6) W-B-B/G-G-W/G Total _____
.....

Training

G-B	W-G	G-W-B	B-W-G-W/G	W/G-G-W-B-B/G	W-B-B/G-G-W/G
G-B	W-G	G-W-B	B-W-G-W/G	W/G-G-W-B-B/G	W-B-B/G-G-W/G
G-B	W-G	G-W-B	B-W-G-W/G	W/G-G-W-B-B/G	W-B-B/G-G-W/G

.....

Administration B

(1) G-B (2) W-G (3) G-W-B (4) B-W-G-W/G (5) W/G-G-W-B-B/G
(6) W-B-B/G-G-W/G Total _____
.....

Administration C

(1) G-B (2) W-G (3) G-W-B (4) B-W-G-W/G (5) W/G-G-W-B-B/G
(6) W-B-B/G-G-W/G Total _____
.....
.....

Administration D

(1) G-B (2) W-G (3) G-W-B (4) B-W-G-W/G (5) W/G-G-W-B-B/G
(6) W-B-B/G-G-W/G Total _____
.....

Administration E

(1) G-B (2) W-G (3) G-W-B (4) B-W-G-W/G (5) W/G-G-W-B-B/G
(6) W-B-B/G-G-W/G Total _____
.....

Comments:

IV. The Light Test



Materials

Two gray boxes, each 14 and 1/2 inches long, 11 inches wide, and 6 and 1/2 inches high. White knobs were used on one of the boxes, and corresponding black knobs on the other. Above the knobs, on each side of the box, was a plastic-covered "window," 6 inches long and 1 inch wide. Behind the windows were a series of flashlight bulbs, which lit when the rectangular knob was pressed. A single rectangle appeared on each side. A rectangle and a sphere were used on Side 1. Side 2 contained a diamond and a rectangle. Side 3 included a rectangle, a diamond, and a

triangle. A diamond, a rectangle, and a triangle, and a sphere were used on Side 4. The spheres were 2 inches in diameter. The triangles were 2 and 1/2 inches on each side. The diamonds were 2 and 3/4 inches long and 1 and 1/2 inches wide. The rectangles were 2 and 1/2 inches high and 1 inch wide.

Pretest Procedures

The box with white knobs was placed about two inches from the edge of the table, Side 1 facing the subject. The examiner pressed the rectangular knob, pointed from the knob to the lighted window, and said, "Light, - see the light." Then the examiner pressed the sphere, pointed from the knob to the unlit window, shook his head, and said, "no light, - see, no light." Next, the examiner pressed the rectangle again, pointing as before, and saying, "Here's the light." Turning to the subject, the examiner said, "You make light," pointing to the subject, and then to the box. If necessary, the examiner pressed the knob at the subject's indication, and corrected the response in case of error. As the rectangle was pressed, the examiner pointed again to the window, repeating, "See the light." The examiner then turned the box so that Side 2 faced the subject, and the entire demonstration was repeated.

Testing Procedures

The test contained eight items, each consisting of the presentation of one side of the box. The order in which the sides were presented is shown on the scoring sheet. As each side was presented, the examiner said, "You make light," pointing from the subject to the box. If necessary, the examiner pressed the knob

which the subject indicated as his choice, and corrected any errors made. All items were given unless the subject failed the first four.

Training Procedures

Training began with Side 1. The examiner pressed the rectangular knob, saying, "here's the light," and pointing from the rectangle to the lighted window, the subject was then asked "to make light," the examiner pointing to the rectangular knob and tapping it until the subject's response was made. The examiner repeated "here's the light" as the rectangular knob was pressed, and corrected errors if necessary.

The subject was trained three times, or through two consecutive successes, on each side of the box. Training was given on each of the four sides, unless the subject achieved no successes in training on the first three sides.

The Transfer Form

The original procedures were followed, using the box with black knobs. The first box was pushed to the back of the testing table, but was not removed from sight.

LIGHT TEST

Administration A

(1) 1 (2) 2 (3) 3 (4) 4 (5) 1 (6) 2 (7) 3 (8) 4 Total _____
.....

Training

Side 1	Side 2	Side 3	Side 4
Side 1	Side 2	Side 3	Side 4
Side 1	Side 2	Side 3	Side 4

.....

Administration B

(1) 1 (2) 2 (3) 3 (4) 4 (5) 1 (6) 2 (7) 3 (8) 4 Total _____
.....

Administration C

(1) 1 (2) 2 (3) 3 (4) 4 (5) 1 (6) 2 (7) 3 (8) 4 Total _____
.....
.....

Administration D

(1) 1 (2) 2 (3) 3 (4) 4 (5) 1 (6) 2 (7) 3 (8) 4 Total _____
.....

Administration E

(1) 1 (2) 2 (3) 3 (4) 4 (5) 1 (6) 2 (7) 3 (8) 4 Total _____
.....

Comments:

V. The Three-Buttons Test



Materials

Three clear plastic boxes, 3 and 1/2 inches high, 3 and 1/2 inches along the bottom, and 4 inches along the top. The boxes were placed 3 inches apart, and attached on parallel sides to two wooden runners, 17 inches long, 1 inch wide, and 1 inch high.

Nine round buttons, 1 and 1/2 inches in diameter, 3 black, 3 white, and 3 gray.

Pretest Procedures

The boxes were placed about two inches from the edge of the table. The examiner dropped a sample white button into the box at the subject's left, saying, "Ping!" as the button struck the bottom of the box. Leaving the sample in place, the examiner placed another white button before the subject, about one inch from the edge of the table and approximately at the center point of the middle box, and said, "You do it," pointing from the subject to the button, and then to the box containing the sample. If necessary, the examiner placed the button at the subject's direction, and corrected or correctly completed the response. These procedures were repeated with black and gray buttons, using the middle and the right hand boxes, respectively. The examiner added "Ping!" as each button was correctly placed. The sample buttons remained in the boxes throughout testing and training.

Testing Procedures

Buttons were presented about an inch from the edge of the table. Single buttons were presented approximately at the center of the boxes. For the two-button items, one button was placed at the mid-point of the space between the left and center boxes, and the other at the corresponding point between the right and center boxes. For the final item, the middle button was placed at the center of the boxes, and the others at the mid-points of the spaces between the center box and those to the left and right, respectively.

The buttons used for each item were removed after its completion. Buttons used in each of the six items are shown on the scoring sheet. "W" represents a white button, "B" a black one,

Al-21

and "0" a gray one. The first three items consisted of placing a single white, black, and gray button into the box containing the corresponding sample, the examiner dropping the buttons into the boxes at the subject's direction, if necessary. The fourth item consisted of a white and a gray button, and the fifth of a gray button and a white one, presented in that order. A black, a gray, and a white button were used for the sixth item. A correct response for the latter items required the correct placement of all of the buttons presented.

As each item was presented, the examiner asked, "Where does this go?" or, "Where do these go?" Pointing from the buttons to the boxes. The examiner accompanied each correct placement with a "Ping!" Errors were corrected, or correctly completed by the examiner. The first three items were always presented. Testing was discontinued if the subject failed the first three items, or the second, third, and fourth. The last item was omitted if the fourth and fifth were failed.

Training Procedures

Training was given three times on each item, or through two consecutive successes. Training always began with the first item. However, if the subject had passed the earlier items on the test, he was trained only once on those items, regular training beginning at the point of his first failure.

Before each item, the examiner demonstrated the correct placement for each button, presented the item, and pointed to the correct box until the subject had responded, saying, "in here," while tapping the edge of the box. All errors were corrected, and

a "Ping!" accompanied a correct response made by the subject, or by the examiner at the subject's indication. Training was discontinued if there was complete failure on the first three items, or on the second, third, and fourth. Training was omitted on the sixth item if the subject failed the fourth and fifth.

The Transfer Form

The samples for the original form of the test were removed, and those used for the transfer situation were placed into the boxes by the examiner, following the procedures used in placing the original samples. The positions of the samples, and the buttons used for each item, are shown on the scoring sheet, (see Administration C). The original testing procedures were followed.

THREE BUTTONS TEST

Administration A

(1) W (2) B (3) G (4) W-G (5) G-W (6) B-G-W Total _____
.....

Training

W	B	G	W-G	G-W	B-G-W
W	B	G	W-G	G-W	B-G-W
W	B	G	W-G	G-W	B-G-W

.....

Administration B

(1) W (2) B (3) G (4) W-G (5) G-W (6) B-G-W Total _____
.....

Administration C

(1) G (2) W (3) B (4) G-B (5) B-G (6) W-B-G Total _____
.....
.....

Administration D

(1) W (2) B (3) G (4) W-G (5) G-W (6) B-G-W Total _____
.....

Administration E

(1) W (2) B (3) G (4) W-G (5) G-W (6) B-G-W Total _____
.....

Comments:

Appendix II

Appendix II

The Experimental Test Battery

This section describes the preliminary battery used in the pilot studies. The various revisions of the tests which were ultimately included in the final battery are described first. The rejected tests are described thereafter.

I. Revisions of Tests Used in the Final Battery.

A. The Two-Buttons Test underwent eight revisions in the pilot studies. The test originally contained five items. The first two consisted of single buttons, one to be placed in the box to the subject's left, and the other in the box to his right. Two, three, and four buttons were then presented, to be dropped into alternate boxes. The resulting scores were too low. The earlier pilot studies therefore concentrated on adding various easy items.

The addition of two single-button items was attempted first. This, however, produced a marked increase in perseveration. When the new items required reversing the order of placement used for the original items, the subject found the shift almost impossible to make, and training merely tended to increase this difficulty rather than to overcome it. On the other hand, when the added items duplicated the original order, the subjects continued using that order regardless of the number of buttons presented. Perseveration decreased to a satisfactory extent when the number of single-button items was reduced to three, and only an alternating order was required for all placements.

Results of various changes in the multiple-button items were fairly consistent. Items containing odd numbers of buttons were found to be more difficult than even-numbered items, to the extent that training for a three-button item tended to be less successful than training for an eight-button item. However, once a subject had grasped the concept of placing buttons in alternating order, increasing the buttons in even numbers presented little additional difficulty. Items consisting of six and eight buttons were therefore dropped as superfluous.

Of the various odd-numbered items attempted, only one three-buttoned item was retained, since inclusion of two items using odd numbers of buttons made the test too difficult. The three-button item was tried out in various positions, and was found to be most useful in next-to-last position.

Various transfer forms for this test were also studied. In some, the examiner used the original white buttons for the demonstrations, presenting black and/or gray buttons to the subject. These were found to be too easy, since higher scores than on the retest administration were obtained. On the other hand, when black and white buttons were both presented to the subject, the resulting scores were too low. Omission of pre-test demonstrations and instructions also made the form too difficult, the subjects apparently failing to understand what was required of them. However, omitting the demonstrations, but pointing to the correct box or boxes in advance of the presentations, seemed to provide sufficient cues.

Of the various sounds tried out as the verbal accompaniment of the demonstrations, "Ping!" was found to be the most interesting to the children, especially when pitched high and given a drawn-out "ng" at the end. Virtually all of the subjects smiled or laughed at this sound. Because of its evident appeal, the sound was reserved as an accompaniment for correct responses only, to increase motivation and reinforce learning.

B. The Buzz Test was revised six times in the preliminary experimentation. As it was originally constructed, the sample discs were arranged on the board in order of size. A round gray peg projected at right angles from the center of each sample, and extending two inches beyond it. Each of the discs presented to the subject had a round hole, 1 and 1/2 inches in diameter, at the center. The discs slipped over the pegs projecting from the samples, the buzz being produced by pressing a disc against its corresponding sample on the board. The majority of the subjects were unable to make the necessary size discriminations for this version of the test. It appeared to be especially difficult for the brain-injured children.

The pegs were removed in the next pilot study, and discs without holes were used. To produce the buzz, the presented disc was laid over the corresponding sample, supported by blunt nails driven into the board just below the samples. This change made the test easier, and did not appear to penalize the brain-injured subjects. However, perseveration, especially after training, was so great that some of the subjects did not shift from the first response for the

remainder of the test. It is interesting to note that perseveration occurred to a significant extent only on this test and the preceding one. It was virtually absent on all other tests throughout their revisions.

In an attempt to overcome perseveration, the differences between the sizes of the discs were increased. This was unsuccessful. In the next attempt, the arrangement of the samples was changed, to avoid a regular increase in size. The smallest disc was placed first and followed by the largest. The next to the smallest came next, with the next to the largest, last. Perseveration was almost entirely overcome. However, the test was now too easy.

To increase the difficulty, presenting each disc twice was attempted. Perseveration increased sharply. Presenting the discs in the order in which they were used in the test, and then repeating the series again in that order, overcame the tendency to perseveration, but made the test too long to hold interest. Two items were therefore dropped. One complete series was given, plus two additional items consisting of one placement at the extreme left and one at the extreme right. This did not give rise to perseveration, and shortened the test sufficiently.

C. The Music Box Test was changed six times. In its original form, it consisted of two gray boxes and two white boxes. The one functioning box was black. Since this form was found to be too easy for purposes of discrimination, two additional boxes were used, one gray and one white. However, this increase in the length of the test tended to make the subjects restless.

The number of boxes was reduced again to five, and one of the gray boxes was darkened considerably. No marked increase in difficulty resulted. However, when the boxes were repainted so that the darker gray box played, the subjects found it too difficult to distinguish between the two shades of gray.

A two-toned box, half white and half black, was introduced in the next attempt, the single gray box playing. Again the test was found to be too easy. Two two-toned boxes, one black and gray and one gray and white were tried next, presented with their gray sides facing the gray box. This tended to confuse the subjects, especially the brain-injured children. In the final form, the two-toned boxes were always presented with their gray sides away from the gray box, which considerably lessened the confusion, and also made the test more suitable for the brain-injured subjects. Since the original rectangular boxes were satisfactory for the transfer form, no changes were made in their size and shape. However, they were painted to correspond with the round boxes in each revision.

D. The Light Test underwent seven revisions. Only three shapes were used in its original construction; spheres, triangles, and rectangles, the latter being the correct response. Three knobs were the maximum used on any one side of the box. Although this form was too easy for the group as a whole, the first two sides were not changed, the simpler items being necessary for discrimination among the more retarded subjects. The revisions therefore concentrated on increasing the difficulty of the remaining items.

A larger box was constructed. The fourth side now containing four knobs, including a second sphere. Since this did not increase the difficulty sufficiently, all spheres were removed from the fourth side, and a diamond was introduced. This was still too easy. One sphere was therefore added, the rectangle being inserted between the sphere and the diamond. Since this also was unsuccessful, the rectangle was placed between the diamond and the triangle, a change which greatly increased the difficulty of the item. The original sphere on the third side of the box was also replaced by a diamond, which increased the difficulty of this item as well.

The Light Test was so interesting to the subjects that the length of the test could be increased without much risk of loss of interest. Two methods were tried for adding to the number of the items. In the first method, each side of the box was presented twice, before the next side was turned to the subject. This produced highly uneven scores, the second item on each side being considerably easier than the first. In the final revision, the sides were presented in the order of difficulty, the entire procedure being repeated twice.

E. The Three-Buttons Test was revised four times in the course of the pilot studies. In its original form in which only single buttons were used, it was found to be too easy for purposes of discrimination. Introducing two-button items did not increase the difficulty sufficiently. On the other hand, proceeding directly from single buttons to three-button items made the test too difficult. In the final form, three single-

button presentations were used, followed by two two-button items, with one three-button item as the final presentation.

II. Rejected Tests.

A. The Sounds Test, a test of shape discrimination, consisted of a gray board like that used in the "Buzz Test," but two inches longer. Projecting from the board were four knobs; an octagon, a triangle, a gear, and a star. Pressing the octagon caused chimes to sound; the triangle rang a bell; the gear produced a buzz. Pressing the star caused no sound. The subject was shown, in pretest demonstrations, the results of pressing the various knobs. He was then asked to reproduce the sound, being instructed "to make ting-a-ling," "to make buzz," and so on. For the star, he was asked "to make sh-sh," the examiner putting a finger to his lips to indicate silence. Various patterns of sounds were used during the pilot studies.

A number of special difficulties arose. One problem was the unwillingness of the subjects to shift from one knob to another. This was apparently not due to a perseverative tendency, since the subjects seemed to be able to shift when they elected to do so. However, they liked the sounds, and, having produced one, they tended to repeat it with evident enjoyment, regardless of instructions.

In an effort to force the shift, cut-off buttons were installed, so that the examiner could shut off the sound after the response was secured. This tended both to confuse and to anger the subjects, who often pressed harder and harder on the

knob, apparently to "make it work." In the next attempt, the examiner demonstrated the next response immediately after completion of the previous one, thus directing the child's attention to the new sound. This was successful in facilitating shifts, but reduced the test to a simple imitation problem.

All attempts to secure responses based on sound patterns were apparently beyond the abilities of the subjects, even when only two sounds were used at a time. The test was finally abandoned as too difficult for purposes of discrimination for this population. It was tried out experimentally with a group of moderately retarded subjects, for whom it appeared to be highly suitable. The subjects enjoyed the test, and were able to handle simple sound patterns. A transfer form based on reversals of the original patterns also appeared to be feasible. Since the test can easily be used for presenting patterns ranging from the very simple to the highly complex, it might serve in an upward extension of this method of testing.

B. The Rattle Test was a test of imitation, in which subject and examiner used comparable sets of material. The subject was required to select the correct material, and use it in imitation of the examiner's demonstration. The materials included a rattle, a small pillow, a hair-brush, and a cup. After each demonstration, the subject was asked to "make rattle - rattle - rattle," "to go to sleep," and so on. A correct response required that the subject indicate the

material to be used, and the part of the body with which its use was associated; i.e., the head for the pillow, the mouth for the cup, and so on.

This test was so appealing to the subjects that it was retained in various forms until the last pilot study. Unfortunately, the revisions did not succeed in making it sufficiently useful for purposes of discrimination. When the materials were presented singly or in groups of twos or threes, the test was too easy. When more complicated patterns were attempted, the test became too difficult. Similar difficulties were found in the various transfer forms. Reversals of the order of the materials was too easy, while the use of the verbal instructions without the demonstrations was too difficult.

Though the seven revisions failed to result in an acceptable form of the test, it is still thought to be potentially useful because of its exceptional appeal to the children. It is also thought to have possibilities for other groups of retarded subjects. Limited experimentation suggested that presentation of the materials singly and in small groups would be suitable for younger severe retardates, while more complicated patterns seemed to be appropriate for the moderately retarded. The test was highly successful in holding the interest of both groups.

C. The Bang Test was also a test of imitation. The examiner struck the table an increasing number of times, saying, "Bang!" at each strike. The subject was then asked "to make bang!" or "to make bang! - bang!" and so on. The same

procedures were repeated for the transfer form, using the vertical surface of a wall instead of the horizontal tabletop. Various items in different orders were tried, the number of strikes ranging from one to seven.

The test presented several problems. A major difficulty was a marked tendency on the part of the subject to respond with random banging, a tendency which training did not often overcome. However, the most serious problem was that a number of the subjects disliked the test. Although they appeared to enjoy the demonstrations, they did not want to imitate them. In the case of subjects with whom the training was successful, another type of difficulty arose. Some continued to strike or bang the material of subsequent tests, disrupting the testing situation.

D and E. The Drum Test and The Tambourine Test were attempts to retain the method of the preceding test. The items in the "Bang Test" were repeated, first by beating against a drum with the hand, and later by striking a tambourine. In both cases, the subjects did enjoy the test. However, their tendency to respond randomly was not lessened, and attempts at training tended to increase it still further.

F. The Bell Test was the most successful of the attempts to utilize the principles of the "Bang Test." A domed white bell was used, in which a clapper hit the metal side of the bell when the projecting top was struck. The bell was very appealing to the subjects, though several revisions were necessary in its construction before it was suitable. A flat white disc

raised well above the bell's dome was used in place of the original top, which was too small to attract the subjects' attention. More successful was a black disc, set high above the dome.

The use of the bell decreased the tendency to respond randomly, although it did not overcome it entirely. However, now that a greater number of correct responses were secured, another problem came into evidence. Subjects who had learned rudimentary counting performed better than subjects who had not. The test was therefore abandoned, since successful performance so evidently depended on differential past experience.

G. The Egg Test was a test of size discrimination. A lure was placed in the largest of a series of wooden eggs of various sizes. This egg was then presented to the subject in a different position for each item, the eggs being used in increasing numbers. Each egg was presented resting on its base, tapered end upward.

Since the first form of the test was found to be too easy, decreasing the differences between the sizes of the eggs was attempted. This change made the test too confusing, especially for the brain-injured subjects. Several attempts were also made to present the egg in a horizontal position. A gray board with oval indentations was used to prevent the eggs from rolling out of position. These attempts were not successful. When the eggs were presented so that the line from base to apex was parallel to the edge of the table, the test was too

easy. On the other hand, the test became too difficult when either the tapered ends of the eggs or their bases were presented toward the subject. Several transfer forms were attempted, using black eggs of the same sizes as the white set, and various other shapes of corresponding relative sizes.

The test was finally abandoned as it became increasingly evident that the "Buzz Test" was a more suitable size discrimination test for this population, and was also easier to administer. However, the "Egg Test" was attempted in its first form with a group of younger severely retarded subjects, for whom the "Buzz Test" was too difficult. Results suggested that the "Egg Test" may be suitable for the younger subjects.